高等院校研究生规划教材

工程数学模型及数值计算方法

刘小华 ◎ 编

石油工业出版社
Petroleum Industry Press

责任编辑：方子奇　谭玉杰

责任校对：罗彩霞

封面设计：翻中天设计

ISBN 978-7-5183-0152-2

9 787518 301522

定价：20.00 元

高等院校研究生规划教材

工程数学模型及数值计算方法

刘小华　编

石油工业出版社

内 容 提 要

本书主要介绍两部分内容：一是数学建模的基本方法、基本步骤和一些常见的数学模型；二是求解数学模型的一些基本数值计算方法，包括插值法、曲线拟合、数值积分、求解方程组的迭代法、方程求根及常微分方程求解的数值方法等. 本书主要结合石油工程中的实际问题引入数值计算方法，实用性强，通俗易懂.

本书可作为全日制专业学位研究生数值计算方法教材，也可作为其他领域的工程技术人员的参考书.

图书在版编目(CIP)数据

工程数学模型及数值计算方法/刘小华编.
北京：石油工业出版社，2014.5
(高等院校研究生规划教材)
ISBN 978-7-5183-0152-2

Ⅰ. 工…
Ⅱ. 刘…
Ⅲ. ①工程数学-数学模型-研究生-教材
②工程数学-数学模型-数值计算-研究生-教材
Ⅳ. TB11

中国版本图书馆 CIP 数据核字(2014)第 072578 号

出版发行：石油工业出版社
(北京安定门外安华里 2 区 1 号 100011)
网 址：http://pip.cnpc.com.cn
编辑部：(010)64523579 发行部：(010)64523620
经 销：全国新华书店
排 版：北京乘设伟业科技有限公司
印 刷：北京中石油彩色印刷有限责任公司

2014 年 5 月第 1 版 2014 年 5 月第 1 次印刷
787×1092 毫米 开本：1/16 印张：10.75
字数：275 千字

定价：20.00 元
(如出现印装质量问题，我社发行部负责调换)

前　言

为更好地适应社会发展对高层次应用型人才的迫切需要，进一步调整和完善硕士研究生教育培养体系，推动硕士研究生教育从以培养学术型人才为主的模式向以培养应用型人才为主的模式转变，我国设立了学术型硕士研究生和全日制专业学位型硕士研究生两种培养模式. 学术型硕士研究生的培养，侧重于理论、学术研究；全日制专业学位型硕士研究生的培养，旨在针对一定的职业背景，培养高层次应用型人才.

经过多年的研究生教育，西南石油大学已积累了丰富的实践经验. 2009 年，教育部确定西南石油大学为全日制专业学位研究生教育试点学校 . 经过几年的努力，已取得初步成效：2011 年招收全日制专业学位研究生 200 人左右，2012 年招收全日制专业学位研究生 400 人左右. 2015 年，全日制专业学位研究生规模将达到学校研究生规模的 50%. 学校指出，全日制专业学位研究生教育是我校研究生教育的发展方向，全校都要重视专业学位研究生的教育，提高培养质量，实现全日制专业学位研究生教育的可持续发展.

搞好全日制专业学位研究生教育，必须做好与之相适应的培养规划及培养方案. 西南石油大学经过多次讨论，已形成了全日制专业学位研究生培养的整体规划并逐步完善. 在全日制专业学位的培养方案中，它的数学课程相比于学术型研究生的数学课程的设置做了较大调整. 我校学术型硕士研究生大多数需要学习“数值分析”和“数学物理方程”两门数学课程，有些同学根据专业需要，还学习了“最优化理论”、“数理统计”或“模糊数学”等课程. 全日制专业学位研究生只学习一门数学课程——“工程数学模型及数值计算方法”，为 60 学时.

《工程数学模型及数值计算方法》是介绍工程中的数学模型以及与数学模型相关的知识，同时介绍求解这些数学模型的数值计算方法或称之为科学计算方法. 这里的“工程”是科学和数学的某种应用，通过这一应用，使自然界物质和能源的特性能够通过各种结构、机器、产品、系统和过程，是以最短的时间和精而少的人力做出高效、可靠且对人类有用的东西. 简言之，工程是将自然科学理论应用到具体工农业生产部门中所形成的各学科的总称，如水利工程、海洋工程、土木工程、化学工程、遗传工程、石油工程、生物工程等. 显然，工程中的数学模型非常之多，数学模型的求解方法或者是数值求解方法也非常多，想要穷尽这些方法是不现实的. 现在已有的与“数学模型”及“数值方法”有关的参考书比比皆是，如何将二者有机地融合在一起并形成一本对全日制专业学位教育有参考价值的教材值得深入研究. 所以编写这样一本书，对我而言简直就是如履薄冰，如临深渊.

经过与研究生院有关领导以及理学院信息与计算科学教研室老师共同探讨，结合我校的实际以及专业学位教育的要求，审慎思考后决定按照“理工结合，涉猎广泛，削弱理论，注重应用”的思路来编写本书. 本书有如下一些特点：

(1)在数学建模方面，注重于让学生理解数学建模的过程，了解数学模型的广泛应用，体会数学建模的艰辛. 涉及的数学模型主要包括常微分方程模型、偏微分方程模型和优化模型. 目的是让学生“欣赏”一些数学模型而并非期望学生能够建立复杂的数学模型.

(2)在数值计算方法方面，重点介绍数值计算方法的思想和原理，不过多分析算法的细节；讨论数值计算方法之间的联系，对方法进行评价.

(3)本书的编写都是强调数学的应用. 数学在越来越定量化研究的今天,地位越来越高,作用越来越大,各行各业需要用到的数学知识也是越来越多. 我校是以石油天然气为特色和优势的学校,石油行业更是技术资金密集的产业,数学在其中凸显出极为重要的作用,对数学的要求也是越来越高. 如果通过本课程的学习,能促使大家进一步认识到数学和工程实际问题的紧密关系,能促进大家进一步应用数学的知识和方法解决实际问题,这也是我的希望所在. 所以在章节的序言部分并没有泛泛而谈该章节知识的重要性,而是首先引入有实际背景的,特别是有石油背景的例子来说明需要讲解的内容.

(4)数学不仅是一门科学,而且是一门技术,数学是可以产生扎扎实实的经济效益的. 美国科学总统顾问委员会前主席戴维说过,高科技本质上是数学技术. 数学也是一门艺术. 体会数学的优雅和美丽,自觉接受数学文化的熏陶,提高自己的数学修养,也是数学教育工作者的责任 . 数学不是高高在上的,也不是深奥难懂的. 本教材注重将数学的严谨性、趣味性与实用性相结合.

(5)由于学时的限制,整个数值计算方法内容部分和一般的数值方法教材相比较,缺少一些内容,比如插值部分,基本上略去有理插值和样条插值的内容,也没有涉及高维插值的计算方法;在曲线拟合部分,略去了函数逼近部分内容;在数值积分部分,略去了高维积分内容;在方程组的迭代法求解部分,没有涉及对称超松弛迭代法;在常微分方程部分,略去了刚性常微分方程的内容,也没有涉及试射法内容(尽管该方法有些陈旧,但有时还是有用的). 另外,本书完全省略了特征值的数值计算方法的介绍 .

笔者认为,研究生阶段的这门课程完全类似于本科生阶段的“高等数学”. 学习了“高等数学”并不能立刻解决工程实际问题,但是如果不学习该课程,其他的课程(包括专业课程)很难学习下去 . 换句话说,本课程是研究生阶段的一门重要的基础课程 . 硕士研究生以及博士研究生在学习研究过程中所遇见的数学问题各种各样,千差万别,涉及的数学分支学科众多 . 单凭本课程或者其他开设的数学课程就能适应学习或研究的需要,这是不符合实际情况的,应注重培养学生提高自身的数学修养,激发学生学习数学的兴趣,鼓励学生在专业学习研究中不断地学习有关数学知识,并应用数学知识解决相关的工程实际问题 .

本书是在 2011 年开始使用的讲义基础上,对将其中的错误进行了修改,并完善和增加了部分内容.

本书由刘小华编写并完成统稿工作,杨艳老师绘制了全书的插图并在编写过程中提出了不少修改建议. 本书的出版,得到了西南石油大学刘志斌教授、谢祥俊教授的大力支持和帮助。长江大学陈忠教授,东北石油大学赵忠奎教授,中国石油大学(北京)崔学慧、张明教授,西安石油大学郝华宁、王忠义教授都提出了不少宝贵建议和意见,在此表示感谢! 同时,感谢西南石油大学教务处的资助! 感谢参考文献的诸多作者!

由于水平有限,教材中错误和不足之处在所难免,恳请读者不吝指正.

刘小华

2014 年 1 月

目　录

1 数学模型基础

近几十年来，数学的应用不仅在它的传统领域——工程技术和经济建设发挥着越来越重要的作用，而且不断地向一些新的领域渗透，形成了许多交叉学科——计量经济学、人口控制论、生物数学、地质数学等．数学与计算机技术相结合，形成了一种普遍的、可以实现的关键技术——数学技术，称为当代高新技术的重要的组成部分．“高技术本质上是数学技术”的观点已被越来越多的人们所接受．

不论是用数学方法解决哪类实际问题，还是与其他学科相结合形成交叉学科，首要的和关键的一步就是用数学的语言表述所研究的对象，即建立数学模型．在高科技，特别是计算机技术迅速发展的今天，计算和建模正成为数学科学技术转化的主要途径．

什么是数学模型？其实大家很早就碰到了数学模型的问题了，如大家以前所熟知的“航行问题”．

甲乙两地相距 750km，船从甲地到乙地顺水航行需 30h，从乙地到甲地逆水航行需要 50h，问船速、水速各是多少？

如果用 x，y 分别表示船速和水速，则有

$$30(x+y)=750,50(x-y)=750.$$

实际上，这组方程就是上述航行问题的数学模型．列出方程，原问题就转化为纯粹的数学问题．方程的解为 $x=20\text{km/h}$、$y=5\text{km/h}$，最终给出了航行问题的答案．

当然，真正实际问题的数学模型通常要复杂得多，但是建立模型的基本内容已经包含在解这个代数应用题的过程中了．那就是：根据建立数学模型的目的和问题的背景做出必要的简化假设（航行中假设船速和水速是常数）；用字母表示待求的未知量；利用相应的物理规律（匀速运动的位移等于速度和时间的乘积）或其他规律，列出相应的数学式子（二元一次方程组）；求出数学上的解（$x=20\text{km/h}$、$y=5\text{km/h}$），用这个答案解释原来问题（船速和水速分别为 20km/h 和 5km/h）；最后还要用实际现象来验证上述结果．

一般地说，数学模型可以描述为，对于现实世界的一个特定对象，为了一个特定目的，根据特有的内在规律，做出一些必要的简化，运用适当的数学工具，得到的一个数学结构．

怎样建立数学模型？数学模型有什么重要意义？这个问题很复杂，下面先看摘自于 2011 年第 7 期《读者》上的一篇文章，希望大家能够细细体会个中深意．

美国普林斯顿大学、麻省理工学院、弗吉尼亚理工学院合作实施一项计划，研究了三年半．研究什么？研究猫俯身伸舌头喝碟子里的牛奶时为什么从来不会弄湿下巴的毛．

三所大学的流体力学家和数学家，都在想这个问题．原来早在 1941 年，美国的科学家就发现猫喝奶时，伸出的舌头就像一个反转过来的字母 J．也就是说，舌头向下卷，把奶很快地捞进嘴巴．

自从发明了高速电子摄影，他们发现猫喝奶的技巧复杂而精妙．猫的舌尖很巧妙地只触及奶的表面，像龙卷风一样，只把奶液向下推，再利用地心引力的反作用力，把奶液像真空管一样吸起来形成一个圆筒形，电光石火之间，送进口腔．

猫的舌头一秒钟可高速舔 4 次，每次喝进 0.1 毫升．科学家把喝奶的纪录片细看，看了三

年半，终于算出了猫舌出击的速度和每次卷舌头的频率之间的一个方程式．再计算猫舌的面积，加进去，就算出了一个“佛罗德函数”的新东西：一只猫每伸一次舌头舔进多少奶，与猫舌面积和伸缩速度的关系．物理学家在前面看片、记录，数学家压阵分析数据，推导出了一个天衣无缝的流体力学新公式．

为什么有此发现？全因为一个叫史托克的流体力学家．一次，他在家里的厨房喂他的宠物猫喝奶，他抚摸着它，欣赏它的美态，忽然兴起研究猫喝奶的念头，像牛顿头上掉下了一只苹果，他无意中发现了神迹．

这是一项研究，也是一种激情，来自对生命的好奇和热爱，对动物的欣赏和呵护．由敏锐的感觉到冷静的深思，成就了西方文明．

这样的研究，中国人觉得无聊——猫喝牛奶有什么好看的？许多父母叫小孩立志做航天员、当总统．他们喜欢大志向，不屑于在小事上钻研、下功夫．由猫喝牛奶，西方的学者研究出大学问来，后面有一股动力——最好的教育，是培养一颗仁心．

《读者》上的这篇文章具有文化娱乐味道，其主旨不外乎是想探究成就西方辉煌文明的一些动因，并以此警戒时下的浮躁之风；至于文中所提及的流体力学公式是不是像文章说的那样研究得到则不必深究．但是从文中的描述也不难看到，建立一个数学模型，特别是建立一个重要的复杂的数学模型的过程是艰辛的，不是一蹴而就的．在本章里，侧重于介绍数学建模的过程，以及数学模型在工程实际问题中的广泛应用．

1.1 数学建模的基本方法和步骤

建立一个数学模型，它面临的实际问题是多种多样的．建模的目的不同、分析的方法不同、采用的数学工具不同，所得到的模型也不同．不能期望归纳出若干条准则，适用于一切实际问题的数学建模方法．下面所谓的基本方法不是针对具体问题而是从方法论的意义上讲的．

1.1.1 数学建模的基本方法

一般来说，建模方法大体上可以分为机理分析和测试分析．机理分析是根据对客观事物特性的认识，找出反映内部机理的数量规律，建立的模型常有明确的物理或现实意义．测试分析将研究对象看做一个“黑箱”系统(内部机理尚不清楚)，通过对系统输入、输出数据的测量和统计分析，按照一定的准则找出数据拟合得最好的模型．

面对一个实际问题用哪种方法建模，主要取决于人们对研究对象的了解和建模的目的．对于许多实际问题还常常将两种方法结合起来进行建模，即用机理分析建立模型的结构，用测试分析确定模型的参数．

1.1.2 数学建模的基本步骤

建模要经过哪些步骤，并没有一定的模式，通常与问题性质、建模目的等有关．下面介绍的是机理分析方法建模的一般过程．

(1)模型准备．了解问题的实际背景，明确建模目的，搜集必要的信息如现象、数据等，尽量弄清楚对象的主要特征，形成一个比较清晰的“问题”，由此初步确定用哪一类模型．在模型准

备阶段要深入调查研究，虚心向实际工作者请教，尽量掌握第一手资料．对实体信息的全面了解，有助于抓住事物的本质，万有引力定律之所以成功，很重要的一点是它的基础工作．第谷布拉赫积累了20年行星运动的数据，为开普勒推导行星三定律奠定了基础，牛顿才有可能发现万有引力定律．因此，建立一个合理的数学模型，详细入微的调研工作是必不可少的．

(2)模型假设. 要用一个抽象的数学结构描述一个复杂的实际问题，必须对问题进行简化．影响问题的因素很多，所以根据问题的特征和建模目的，抓住问题的本质，忽略次要因素，做出必要的合理的假设．对于建模的成败这是非常重要的和困难的一步．不同的假设会得到不同的数学模型，人口模型就是典型的例子．自Malthus的第一个人口模型问世至今，已有一百多个人口模型相继产生．建模者依据不同的假设，建立不同的模型，从不同的角度论述人口问题，这也表明恰当的假设和建模的目的密切相关．另外，假设做得不合理或太简单，会导致错误的或无用的模型；假设做得过分详细，试图把复杂对象的众多因素都考虑进去，会很难或无法继续下一步的工作．常常需要在合理与简化之间做出恰当的折中．通常，做假设的依据，一是出于对问题内在规律的认识，二是来自对现象、数据的分析，以及二者的综合．想象力、观察力、判断力，以及经验，在模型假设都起着非常重要的作用．

(3)模型构成. 根据所作的假设，用数学语言、符号描述对象的内在规律，建立包含变量、常量等的数学模型，如优化模型、微分方程模型、差分方程模型等．这里除了需要相关学科的专门知识外，还常常需要较为广阔的应用数学知识．要善于发挥想象力，注意使用类比法，分析对象与熟悉的其他对象的共性，借用已有的模型．

(4)模型求解. 可以采用解方程、画图形、优化方法、数值计算、统计分析等各种数学方法，特别是数学软件和计算机技术．

(5)模型分析. 对求解结果进行数学上的分析，如结果的误差分析、统计分析、模型对数据的灵敏性分析等．

(6)模型检验. 把求解和分析结果翻译回到实际问题，与实际的现象、数据比较，检验模型的合理性和适用性．如果与实际不符，问题常常出在模型假设上，应该修改、补充假设，重新建模．这一步对于模型是否真的有用非常关键，要以严肃认真的态度对待，有些模型要经过几次反复，不断完善，直到检验结果获得某种程度上的满意．例如，牛顿创立的万有引力定律就经受了对哈雷彗星的研究、海王星的发现等大量事实的考验，最终被公认为是一个成功的数学模型．

1.2 数学模型的特点与分类

1.2.1 数学模型的特点

数学建模是利用数学工具解决实际问题的重要手段，得到的模型有许多优点，也有一些缺点．下面归纳出数学模型的若干特点．

(1)模型的逼真性和可行性. 一般来说总是希望模型将尽可能逼近研究对象，但是一个非常逼真的模型在数学上常常是难于处理的，因而不容易达到通过建模对现实对象进行分析、预报、决策或者控制的目的，即实用上不可行．另一方面，越逼真的模型常常越复杂，即使数学上能处理，这样的模型在应用时所需要的“费用”也相当高，而高“费用”不一定与复杂模型取得的

“效益”相匹配．所以，建模时往往需要在模型的逼真性与可行性、“费用”与“效益”之间做出权衡．

(2)模型的渐进性．稍微复杂一些的实际问题的建模通常不可能一次成功，要经过反反复复的过程，包括由简到繁、删繁就简，以获得越来越满意的模型．在科学发展中，随着人们认识和实践能力的提高，各门学科中的数学模型也存在一个不断完善或者推陈出新的过程．从19世纪力学、热学、电学等许多学科由牛顿力学的模型主宰，到20世纪爱因斯坦相对论模型的建立，这些都是模型渐进性的明显例证．

(3)模型的强健性．模型的结构和参数常常是模型假设及对象的信息如观测数据确定的，而假设可能不太准确，观测数据可能也允许有误差．一个好的模型应该具有下述意义的强健性：当模型假设改变时，可以导出模型结构的相应变化；当观测数据有微小变化时，模型的参数也只有相应的微小变化．

(4)模型的可转移性．模型是现实对象抽象化、理想化的产物，它不为对象的所属领域所独有，可以转移到另外的领域．在生态、经济、社会等领域内建模就经常借用物理学领域的模型．模型的这种性质显示了它应用的极端广泛性．

(5)模型的非预制性．虽然已经发展了许多应用广泛的模型，但是实际问题是各种各样、变化万千的，不可能要求把各种模型做成预制品供你在建模时使用．模型的这种非预制性使得建模本身常常是事先没有答案的问题(Open－end problem)．在建立新的模型的过程中甚至会伴随新的数学方法或数学概念的产生．

(6)模型的技艺性．建模的方法与其他一些数学方法如方程解法、规划问题解法等根本不同，无法归纳出若干条普遍适用的建模准则和技巧．建模目前与其说是一门技术，不如说是一门艺术，是技艺性很强的技巧．经验、想象力、洞察力、判断力，以及直觉、灵感等在建模过程中起的作用往往比一些具体的数学知识更大．

(7)模型的局限性．第一，由数学模型得到的结论虽然具有通用性和精确性，但是因为模型是对现实对象的简化、理想化的产物，所以一旦将模型的结论应用于实际问题，就回到了现实世界，那些被忽视、简化的因素必须考虑，于是结论的通用性和精确性只是相对的近似．第二，由于人们认识能力和科学技术包括数学本身发展水平的限制，还有不少实际问题很难得到有实际价值的数学模型．如一些内部机理复杂、影响因素众多、测量手段不够完善、技艺性较强的生产过程，像生铁冶炼过程，常常需要开发专家系统，与建立数学模型相结合才能获得较满意的应用效果．专家系统是一种计算机系统，它总结专家的知识和经验，模拟人类的逻辑思维过程，建立若干规则和推理途径，主要是定性地分析各种实际现象并作出判断．专家系统可以看成是计算机模拟的新发展．第三，还有一些领域中的问题至今尚未发展到用建模方法寻求数量规律的阶段，如中医诊断过程，目前所谓的计算机辅助诊断也是属于总结著名中医的丰富临床经验的专家系统．

1.2.2 数学模型的分类

数学模型可以按照不同的方式分类，下面介绍常用的几种．

(1)按照模型的应用领域分类，如人口模型、交通模型、环境模型、生态模型等．范畴更大一些则形成许多边缘学科，如生物数学、医学数学、地质数学、数量经济学等．

(2)按照建立模型的数学方法(或所属数学分支)分类，如初等模型、几何模型、微分方程模型、统计回归模型、数学规划模型等．

(3)按照模型的表现特性又有以下几种分类法：

确定性模型和随机性模型——取决于是否考虑随机因素的影响．近年来随着数学的发展，又有所谓突变型模型和模糊行模型．

静态模型和动态模型——取决于是否考虑时间因素引起的变化．

线性模型和非线性模型——取决于模型的基本关系，如微分方程是否是线性的．

离散模型和连续模型——取决于模型中的变量(主要是时间变量)取为离散还是连续的．

虽然从本质上讲大多数实际问题是随机的、动态的、非线性的，但是由于确定性、静态、线性模型容易处理，并且往往可以作为初步的近似来解决问题，所以建模时常先考虑确定性、静态、线性模型．连续模型便于利用微积分方法求解析解，作理论分析，而离散模型便于在计算机上作数值计算，所以用哪种模型要视具体问题而定．在具体的建模过程中将连续问题离散化，或将离散变量视作连续的，也是常用的方法．

(4)按照建模的目的分类，有描述模型、预报模型、优化模型、决策模型、控制模型等．

(5)按照对模型结构的了解程度分类，有所谓白箱模型、灰箱模型、黑箱模型．这是把研究对象比喻成一只箱子里的机关，要通过建模来解释它的奥妙．白箱主要包括用力学、热学、电学等一些机理相当清楚的学科描述的现象及相应的工程技术问题，这方面的模型大多已经基本确定．灰箱主要指生态、气象、经济、交通等领域中机理尚不清楚的现象，在建立和改善模型方面还不同程度地有许多工作要做．至于黑箱则主要指生命科学和社会科学等领域中的一些机理(数量关系方面)很不清楚的现象．有些工程技术问题虽然主要基于物理、化学原理，但是由于因素众多、关系复杂和观测困难等原因也常作为灰箱或黑箱模型处理．

1.3 数学模型实例

1.3.1 赝品的鉴定

1945年第二次世界大战后，荷兰安全机关以通敌罪逮捕了一名三流画家H. A. Vanmeergren. 此人曾将17世纪荷兰著名画家Jan Vermeer创作的一批名贵油画倒卖给德国人．可是Vanmeegren被捕后不久即在狱中宣称，他从未出卖过荷兰的利益，所有油画均是他自己伪造的，这件事在当时轰动了全球．为了证明自己是一个高明的伪造者，他开始在牢房作画．当这幅画快要完工时，他得悉自己可能会改判为伪造罪．为了逃避判决，他未将此画画完，并拒绝将画老化，以免留下罪证．

在审理这一案件时，法庭组织了一个由著名化学家、物理学家、艺术史学家等多学科的专家参加的国际小组．科学家们采用了当时最为先进的科学方法，例如用X光射线透视，用化学分析方法分析颜料的成分等，终于在其中几幅画中发现了20世纪才可能有的某些有机化合物，判定这几幅画确系伪造，并判了他一年徒刑(此人后来因心脏病发作死于狱中)．

如何判断其余的画究竟是不是赝品？这一问题悬而未决．直到1967年，Carneigie－Mellon大学的科学家们根据放射性原理，建立了与此问题相关的一个简单微分方程，利用测得的一些数据，基本上解决了这一问题．

下面仅从判定画作“EMMAUS的信徒”是否属于伪造来大致说明这一建模过程．所有绘画中都含有放射性铅－210(^{210}Pb)和镭－226(^{226}Ra)．这两种元素存在于铅白中，画家们用铅

白做颜料.

设 $y(t)$是铅白中^{210}Pb 的含量,$r(t)$是 t 时刻每克铅白中^{226}Ra 的衰变数量,则 $y(t)$满足常微分方程

$$\begin{cases}\dfrac{dy}{dt}=-\lambda y+r(t),\\ y(t_0)=y_0.\end{cases}\tag{1.1}$$

由于^{226}Ra 的半衰期为 1600 年,而现在仅对 300 年左右的时间感兴趣,所以可以假设 $r(t)$等于常数 r. 容易求出上述方程的解为

$$y=\frac{r}{\lambda}[1-e^{-\lambda(t-t_0)}]+y_0e^{-\lambda(t-t_0)}.\tag{1.2}$$

$y(t)$和 r 可以用仪器测量,为了求 $t-t_0$,只需要想办法求出 y_0 与 λ.

下面计算 λ. 著名物理学家卢瑟福指出:物质的放射性正比于物质的原子数,即若以 $N(t)$表示 t 时刻放射性物质的原子数,则有

$$\begin{cases}\dfrac{dN}{dt}=-\lambda N,\\ N(t_0)=N_0,\end{cases}$$

从而有

$$N(t)=N_0e^{-\lambda(t-t_0)}.$$

已知铅的半衰期为 22 年,由此得到

$$\lambda=\frac{\ln 2}{22}.$$

再计算 y_0. 由前面式(1.2)可得

$$\lambda y_0=\lambda y(t)e^{\lambda(t-t_0)}-r[e^{\lambda(t-t_0)}-1].\tag{1.3}$$

如果画是真品,已有 300 年的历史,则上式中 $t-t_0=300$,带入式中可得

$$\lambda y_0=\lambda y(t)e^{300\lambda}-r(e^{300\lambda}-1).$$

镭^{226}Ra 的衰变率 $r=0.8$,铅^{210}Pb 的衰变率 $\lambda y=8.5$,所以

$$\lambda y_0=8.5e^{300\lambda}-0.8(e^{300\lambda}-1)=98050.$$

而由其余的知识容易推算该值不能大于 30000. 从而可以断定该画“Emmaus 的信徒”是伪造的.

1.3.2 人口增长的预测

长期以来,人类的繁殖一直在自发进行着. 只是由于人口数量的迅速膨胀和环境质量的急剧恶化,人们才猛然醒悟,开始研究人类和自然的关系、人口数量的变化规律,以及如何进行人口控制等问题.

我国是世界第一人口大国，地球中每九个人中就有一个中国人．在20世纪的一段时间内，我国人口的增长速度过快，见表1.1.

表1.1　我国人口数据

年份	1908	1933	1953	1964	1982	1990	2000
人口(亿)	3.0	4.7	6.0	7.2	10.3	11.3	12.95

认识人口数量的变化规律，建立人口模型，做出准确的预测，是有效控制人口增长的前提．长期以来人们在这方面做了不少工作．下面介绍两个最基本的人口模型，并利用表1.2给出的近两个世纪的美国人口统计数据(以百万为单位)对模型进行检验，最后用它预报2010年美国人口．

表1.2　美国人口数据

年份	人口(百万)	年份	人口(百万)	年份	人口(百万)
1790	3.9	1870	38.6	1950	150.7
1800	5.3	1880	50.2	1960	179.3
1810	7.2	1890	62.9	1970	204.0
1820	9.6	1900	76.0	1980	226.5
1830	12.9	1910	92.0	1990	251.4
1840	17.1	1920	106.5	2000	281.4
1850	23.2	1930	123.2		
1860	31.4	1940	131.7		

两百多年前，英国人口学家马尔萨斯(Malthus，1766—1834)调查了英国一百多年的人口统计资料，得出了人口增长率不变的假设，并据此建立了著名的人口指数增长模型．

1.3.2.1　指数增长模型

1)模型建立

记时刻 t 的人口为 $x(t)$，当考察一个国家或一个较大地区的人口时，$x(t)$ 是一个很大的整数．为了利用微积分这一数学工具，将 $x(t)$ 视为连续可微函数．记初始时刻($t=0$)的人口为 x_0．假设人口增长率为常数 r，即单位时间内 $x(t)$ 的增量等于 $rx(t)$．考虑 t 到 $t+\Delta t$ 时间内人口的增量，显然有

$$x(t+\Delta t)-x(t)=rx(t)\Delta t.$$

令 $\Delta t\to 0$，得到 $x(t)$ 满足的微分方程

$$\frac{\mathrm{d}x}{\mathrm{d}t}=rx,x(0)=x_0, \tag{1.4}$$

易见

$$x(t)=x_0\mathrm{e}^{rt}. \tag{1.5}$$

$r>0$ 时，式(1.5)表示人口将按照指数规律随时间无限增长，称为指数增长模型．马尔萨斯因此悲观认为人口增长最终会导致瘟疫、灾祸和饥荒．

2)参数估计

式(1.5)中的参数 r 和 x_0 可以用表1.2中的数据进行估计．为了利用最简单的线性最小

二乘法，将式(1.5)取对数可得

$$y = rt + a, y = \ln x, a = \ln x_0. \tag{1.6}$$

以1790—1900年的数据拟合式(1.6)可得 $r=0.2743/10$ 年，$x_0=4.1884$. 若以全部数据(1790—2000年)进行拟合式(1.6)可得 $r=0.2022/10$ 年，$x_0=6.0450$.

3)结果分析

用上面得到的参数 r 和 x_0 带入式(1.5)，将计算结果与实际数据作比较．表1.3中计算人口 x_1 是用1790—1900年的数据拟合的结果，计算人口 x_2 是用1790—2000年的数据拟合的结果．可以看出，用这个模型基本上能够描述19世纪以前美国人口的增长，但是进入20世纪后，美国的人口增长明显变缓，这个模型就不适合了．

表1.3　人口数据与计算人口数据

年份	实际人口（百万）	计算人口 x_1	计算人口 x_2	年份	实际人口（百万）	计算人口 x_1	计算人口 x_2	年份	实际人口（百万）	计算人口 x_1	计算人口 x_2
1790	3.9	4.2	6.0	1870	38.6	37.6	30.5	1950	150.7		153.6
1800	5.3	5.5	7.4	1880	50.2	49.5	37.3	1960	179.3		188.0
1810	7.2	7.2	9.1	1890	62.9	65.1	45.7	1970	204.0		230.1
1820	9.6	9.5	11.1	1900	76.0	85.6	55.9	1980	226.5		281.7
1830	12.9	12.5	13.6	1910	92.0		68.4	1990	251.4		344.8
1840	17.1	16.5	16.60	1920	106.5		83.7	2000	281.4		422.1
1850	23.2	21.7	20.30	1930	123.2		102.5				
1860	31.4	28.6	24.90	1940	131.7		125.5				

历史上，指数增长模型与19世纪以前欧洲一些地区人口统计数据可以很好的吻合，迁往加拿大的欧洲移民后代人口也大致符合这个模型．另外，用它做短期人口预测可以得到较好的结果．显然，这是因为在这些情况下，模型的基本假设——人口增长率是常数大致成立．

但是从长期来看，任何地区的人口都不能无限增长，即指数增长模型不能描述也不能预测较长时间的人口演变过程．这是因为，人口增长率事实上是不断变化的．排除灾难、战争等特殊时期，一般来说，当人口较少时增长较快，即增长率较大；人口增长到一定数量以后，增长就会慢下来，即增长率变小．表1.4是用数值微分的三点公式计算美国人口增长率(%年)，可以看到，进入20世纪后增长率明显下降．用平均增长率作为 r，用指数增长模型描述美国人口的变化，结果自然和表1.2的统计数据相差很大．

表1.4　计算人口增长率

年份	增长率(%)	年份	增长率(%)	年份	增长率(%)
1790	2.95	1870	2.44	1950	1.58
1800	3.11	1880	2.42	1960	1.49
1810	2.99	1890	2.05	1970	1.16
1820	2.97	1900	1.91	1980	1.05
1830	2.91	1910	1.66	1990	1.09
1840	3.01	1920	1.46	2000	1.16
1850	3.08	1930	1.02		
1860	2.45	1940	1.04		

1.3.2.2 阻滞增长模型(Logistic 模型).

1)模型建立

分析人口增长到一定数量后增长率下降的主要原因,人们注意到,自然资源、环境条件等因素对人口的增长起着阻滞作用,并且随人口的增加,阻滞作用越来越大. 所谓阻滞增长模型就是考虑这个因素,对指数增长模型的基本假设进行修改后得到的.

阻滞作用体现在对人口增长率 r 的影响上,使得 r 随着人口数量 x 的增加而下降. 若将 r 表示为 x 的函数 $r(x)$,则它应是减函数. 于是式(1.4)写作

$$\frac{\mathrm{d}x}{\mathrm{d}t} = r(x)x, x(0) = x_0. \tag{1.7}$$

对 $r(t)$的一个最简单的假定是,设 $r(x)$为 x 的线性函数,即

$$r(x) = r - sx \qquad (r > 0, s > 0). \tag{1.8}$$

这里 r 称为固有增长率,表示人口很少时(理论上是 $x=0$ 的增长率). 为了确定系数 s 的意义,引入自然资源和环境条件所能容纳的最大人口数量 x_m,称为人口容量. 当 $x=x_m$ 时人口不再增长,即增长率 $r(x_m)=0$,带入式(1.8)得 $s=\dfrac{r}{x_m}$,于是式(1.8)为

$$r(x) = r\left(1 - \frac{x}{x_m}\right), \tag{1.9}$$

其中参数 r、x_m 可以通过数据表拟合得到,或者通过人口专家经验确定.

将式(1.9)代入式(1.7)得到

$$\frac{\mathrm{d}x}{\mathrm{d}t} = rx\left(1 - \frac{x}{x_m}\right), x(0) = x_0. \tag{1.10}$$

由分离变量法可知

$$x(t) = \frac{x_m}{1 + \left(\dfrac{x_m}{x_0} - 1\right)\mathrm{e}^{-rt}}. \tag{1.11}$$

用拟合的方式得到参数 $r=0.2557/10$ 年,$x_m=392.0886$. 这时将计算结果和实际数据作比较得到表 1.5.

表 1.5 人口数据与计算人口数据

年份	实际人口(百万)	计算人口 x	年份	实际人口(百万)	计算人口 x	年份	实际人口(百万)	计算人口 x
1790	3.9	3.9	1860	31.4	22.3	1930	123.2	103.9
1800	5.3	5.0	1870	38.6	28.3	1940	131.7	124.5
1810	7.2	6.5	1880	50.2	35.8	1950	150.7	147.2
1820	9.6	8.3	1890	62.9	45	1960	179.3	171.3
1830	12.9	10.7	1900	76.0	56.2	1970	204.0	196.2
1840	17.1	13.7	1910	92.0	69.7	1980	226.5	221.2
1850	23.2	17.5	1920	106.5	85.5	1990	251.4	245.3

从表 1.5 可以看出，这个模型拟合时虽然中间一段(19 世纪中叶到 20 世纪中叶)不大好，但是最后一段(20 世纪中叶以后)吻合得不错.

2)模型检验

在估计阻滞增长模型的参数时没有用到 2000 年的数据，是为了用它作模型检验. 模型的计算结果是 $x(2000)=274.5$ 百万，与实际数据 281.4 百万相比误差约为 2.5%.

3)人口预测

将 2000 年的实际数据加进去重新估计参数可得 $r=0.2490/10$ 年，$x_m=433.9886$. 然后再预报美国 2010 年的人口得到 $x(2010)=306.0$ 百万. 2010 年美国实际人口为 $x(2010)=308.7$ 百万，与实际数据相比误差约为 0.87%.

式(1.10)表示的人口阻滞增长模型，是荷兰生物学家 Verhulst 在 1838 年提出的. 它不仅能够大体上描述人口及许多物种数量(如森林中的树木、鱼塘中的鱼群等)的变化规律，而且在社会经济领域中也有广泛的应用.

在 20 世纪初，由于概率论的发展，G. V. Yule 在 1924 年引入概率论的观点，建立了 Yule 模型. 1945 年 P. H. Leslie 给出人口离散化的人口模型，Leslie 矩阵成为研究种群发展的重要工具. 在 20 世纪中期，由于控制论的迅速发展，Van. H. Fpoerster 在 1959 年提出了连续人口发展模型. 此后，国内外学者对人口问题的研究一直不断深入进行，人口模型也越来越精细. 迄今为止，已有几百个人口模型问世.

1.3.3 压裂液滤失系数模型

水力压裂(Hydraulic fracturing)是人们利用地面高压泵组，以超过地层吸收能力的排量将高黏压裂液泵入井内而在井内憋起高压，当该压力克服井壁附近地应力并达到岩石抗张强度时，就在地层产生裂缝. 继续注入带有支撑剂的混砂液，使裂缝继续延伸并在其中充填支撑剂. 停泵后，由于支撑剂对裂缝的支撑作用，在地层中形成足够长的、有一定导流能力的填砂裂缝，从而实现油气井增产和注水井增注.

在压裂中，压裂液沿裂缝向地层滤失. 一般认为，这种滤失受三种机理的控制——滤液黏度、地层流体的压缩性和压裂液的造壁性. 压裂液滤失量的多少常以滤失系数的大小来表示.

在计算受地层流体压缩性控制的滤失系数时，可导出下面的方程：

$$\begin{cases} \dfrac{\partial^2 p}{\partial x^2}=\dfrac{1}{\eta}\dfrac{\partial p}{\partial t}, \\ p(x,0)=p_i, \\ p(0,t)=p_i+\Delta p, \\ p(\infty,t)=p_i. \end{cases}$$

式中 $p(x,t)$——点 x 在时间 t 时的压力，MPa；

p_i——原始地层压力，MPa；

η——地层导压系数，m^2/s.

利用拉普拉斯变换可以求得该问题的精确解是：

$$p(x,t)=\operatorname{erfc}\left(\frac{x}{2\sqrt{\eta t}}\right)\Delta p+p_i$$

这里 erfc(x)表示余误差函数，

$$\operatorname{erfc}(x)=1-\operatorname{erf}(x)=1-\frac{2}{\sqrt{\pi}}\int_0^x \mathrm{e}^{-t^2}\,\mathrm{d}t=\frac{2}{\sqrt{\pi}}\int_x^{+\infty}\mathrm{e}^{-t^2}\,\mathrm{d}t$$

1.3.4 单相渗流的连续性方程

油气渗流过程必须遵守质量守恒定律(又称连续性原理). 这一定理一般可以描述为:在地层中任取一个微小的单元体,在单元体内若没有源和汇存在,那么包含在单元封闭表面之内的流体质量变化应等于同一时间间隔内流体流入质量和流出质量之差. 用质量守恒原理建立起来的方程称为质量守恒方程,又称为连续性方程.

连续性方程的表现形式是给出运动要素(速度、密度、饱和度、浓度)随时间和坐标的变化关系.

下面利用微元分析法来建立质量守恒方程. 在地层中取微小六面体单元(图 1.1),单元体中 $M(x,y,z)$点质量速度在各坐标轴上分量为 ρv_x、ρv_y 和 ρv_z.

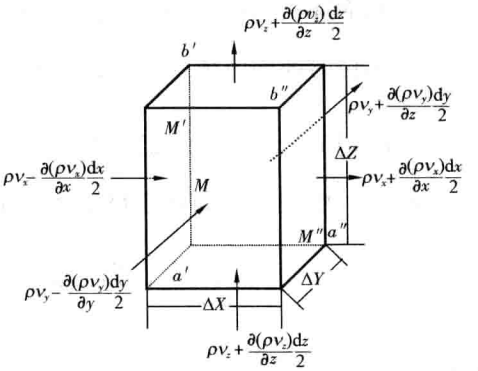

图 1.1 长方体微元

注意到点 $M'\left(x-\frac{\mathrm{d}x}{2},y,z\right)$、点 $M''\left(x+\frac{\mathrm{d}x}{2},y,z\right)$,故点 M'的质量速度为

$$\rho v_x-\frac{\partial(\rho v_x)}{\partial x}\frac{\mathrm{d}x}{2},$$

点 M''的质量速度为

$$\rho v_x+\frac{\partial(\rho v_x)}{\partial x}\frac{\mathrm{d}x}{2}.$$

从而在 $\mathrm{d}t$ 时间内,沿 x 方向流入和流出单元体内的流体质量分别为

$$\left[\rho v_x-\frac{\partial(\rho v_x)}{\partial x}\frac{\mathrm{d}x}{2}\right]\mathrm{d}y\mathrm{d}z\mathrm{d}t$$

和

$$\left[\rho v_x+\frac{\partial(\rho v_x)}{\partial x}\frac{\mathrm{d}x}{2}\right]\mathrm{d}y\mathrm{d}z\mathrm{d}t.$$

从而在 $\mathrm{d}t$ 时间内,沿 x 方向流入和流出的流体质量差为

$$-\frac{\partial(\rho v_x)}{\partial x}\mathrm{d}x\mathrm{d}y\mathrm{d}z\mathrm{d}t.$$

同理在 $\mathrm{d}t$ 时间内,沿 y、z 方向流入和流出的流体质量差为

$$-\frac{\partial(\rho v_x)}{\partial y}\mathrm{d}x\mathrm{d}y\mathrm{d}z\mathrm{d}t$$

和

$$-\frac{\partial(\rho v_z)}{\partial z}\mathrm{d}x\mathrm{d}y\mathrm{d}z\mathrm{d}t.$$

在 dt 时间内，六面体内流入和流出的总的流体质量差为

$$-\left[\frac{\partial(\rho v_x)}{\partial x}+\frac{\partial(\rho v_y)}{\partial y}+\frac{\partial(\rho v_z)}{\partial z}\right]\mathrm{d}x\mathrm{d}y\mathrm{d}z\mathrm{d}t.$$

经过六面体流入和流出的质量之所以会不一样，是因为在六面体内岩石和液体弹性能量的作用下，释放或储存一部分质量的结果(岩石的弹性表现为孔隙度的变化，液体弹性表现为液体密度的变化). 六面体内的质量变化计算如下：

(1)六面体内的孔隙体积为 $\phi\mathrm{d}x\mathrm{d}y\mathrm{d}z$，

(2)六面体内的流体质量为 $\rho\phi\mathrm{d}x\mathrm{d}y\mathrm{d}z$，

(3)单位时间内流体质量变化率为 $\dfrac{\partial(\rho\varphi)}{\partial t}\mathrm{d}x\mathrm{d}y\mathrm{d}z$，

(4)dt 时间内流体质量总的变化为 $\dfrac{\partial(\rho\varphi)}{\partial t}\mathrm{d}x\mathrm{d}y\mathrm{d}z\mathrm{d}t$.

显然，dt 时间内六面体质量总的变化应等于六面体在 dt 时间内流入和流出的质量差，即

$$-\left[\frac{\partial(\rho v_x)}{\partial x}+\frac{\partial(\rho v_y)}{\partial y}+\frac{\partial(\rho v_z)}{\partial z}\right]\mathrm{d}x\mathrm{d}y\mathrm{d}z\mathrm{d}t=\frac{\partial(\rho\varphi)}{\partial t}\mathrm{d}x\mathrm{d}y\mathrm{d}z\mathrm{d}t,$$

消去 $\mathrm{d}x\mathrm{d}y\mathrm{d}z\mathrm{d}t$，得到

$$-\left[\frac{\partial(\rho v_x)}{\partial x}+\frac{\partial(\rho v_y)}{\partial y}+\frac{\partial(\rho v_z)}{\partial z}\right]=\frac{\partial(\rho\varphi)}{\partial t}.$$

上式还可以写成

$$\frac{\partial(\rho\varphi)}{\partial t}+\mathrm{div}(\rho\boldsymbol{v})=0,$$

上式就是单相均质可压缩流体在弹性孔隙介质中的质量守恒方程.

如果再利用流体的运动方程以及岩石和流体的状态方程可以演算得到如下线性三维抛物型方程

$$\frac{\partial^2 p}{\partial x^2}+\frac{\partial^2 p}{\partial y^2}+\frac{\partial^2 p}{\partial z^2}=\frac{1}{\eta}\frac{\partial p}{\partial t},$$

上式就是弹性孔隙介质单相可压缩液体的渗流数学模型，其中 η 表示导压系数.

除了单相渗流之外，还可以考虑油水两相渗流和油气两相渗流问题，在不考虑毛管压力的情况下，和上面类似的方法可以导出两相渗流的质量守恒方程

$$-\left(\frac{\partial v_{\mathrm{ox}}}{\partial x}+\frac{\partial v_{\mathrm{oy}}}{\partial y}+\frac{\partial v_{\mathrm{oz}}}{\partial z}\right)=\varphi\frac{\partial S_{\mathrm{o}}}{\partial t}(\text{对油相}),$$

$$-\left(\frac{\partial v_{\mathrm{wx}}}{\partial x}+\frac{\partial v_{\mathrm{wy}}}{\partial y}+\frac{\partial v_{\mathrm{wz}}}{\partial z}\right)=\varphi\frac{\partial S_{\mathrm{w}}}{\partial t}(\text{对水相}).$$

如果再利用两相的运动方程

$$\boldsymbol{v}_{\mathrm{o}}=-\frac{K_{\mathrm{o}}(S)}{\mu_{\mathrm{o}}}\mathrm{grad}\,p(\text{对油相}),$$

$$v_w = -\frac{K_w(S)}{\mu_w}\text{grad}p(\text{对水相}),$$

可以得到

$$\frac{\partial}{\partial x}\left[\frac{K_o(S)}{\mu_o}\frac{\partial p}{\partial x}\right]+\frac{\partial}{\partial y}\left[\frac{K_o(S)}{\mu_o}\frac{\partial p}{\partial y}\right]+\frac{\partial}{\partial z}\left[\frac{K_o(S)}{\mu_o}\frac{\partial p}{\partial z}\right]=\varphi\frac{\partial S_o}{\partial t},$$

$$\frac{\partial}{\partial x}\left[\frac{K_w(S)}{\mu_w}\frac{\partial p}{\partial x}\right]+\frac{\partial}{\partial y}\left[\frac{K_w(S)}{\mu_w}\frac{\partial p}{\partial y}\right]+\frac{\partial}{\partial z}\left[\frac{K_w(S)}{\mu_w}\frac{\partial p}{\partial z}\right]=\varphi\frac{\partial S_w}{\partial t}.$$

其中，$K_o(S)$和$K_w(S)$分别表示油和水的相渗透率，S是饱和度．

如果考虑毛管压力，将会得到一个复杂的二阶非线性偏微分方程，这种方程只有在某些简单情况下才有精确解．

1.3.5 优化模型

日常生活与科学研究中经常出现最优化问题，比如有多种工作要安排，先做哪些工作，后做哪些工作？每种工作人力、物力、技术手段如何使用？每种工作进行的时间如何？要达到什么样的指标？安排在什么位置进行等，都要做具体安排．安排得好，效率就高，否则，就要窝工．例如，制造一种产品，用什么原料？采用什么规格？什么工艺？什么工序？什么时间生产？生产多少？等等．采用不同方式对产品质量、产值都有影响．一座建筑或一个结构，当主要的要求确定之后，用什么结构形式？什么材料？什么规格？在保证满足相同技术要求情况下，方案不同，投资则会有很大区别．某一物资，产地和销地都很多，安排不同运输方案，运费也将有很大变化．城市、工厂、农村总的平面布局，以及各个小的单位的平面布局等都有很多不同方案，用不同方案就会有不同效果．又如，企业和管理、经济发展的规划、军事行动的指挥等等．

为解决这些问题，不仅形成了最优化理论，而且与之配套形成了一个新的学科分支——最优化技术．这一技术的本质就是将实际问题抽象成数学模型，并用最优化的有关理论求解其数学模型．

从数学上看，定量的最优化问题就是寻找n元函数$f(\boldsymbol{X})$的极值点．当f是普通函数，$\boldsymbol{X}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}\in\mathbf{R}^n$，这类优化问题称为数学规划，变量$\boldsymbol{X}$可能没有限制，也可能受有限个等式或不等式约束．其一般模型可表示成

$$\begin{aligned}(\text{MP})\qquad &\min f(\boldsymbol{X}) \quad (\max f(\boldsymbol{X}))\\ &g_i(\boldsymbol{X})\geqslant 0 \quad i=1,2,\cdots,m\\ &h_j(\boldsymbol{X})=0 \quad j=1,2,\cdots,p\end{aligned}$$

其中$x_i(i=1,2,\cdots,n)$称为决策变量，$f(\boldsymbol{X})$称为目标函数．$g_i(\boldsymbol{X})\geqslant 0\quad(i=1,2,\cdots,m)$，$H_j(x)=0\quad(j=1,2,\cdots,p)$称为约束条件，$g_i(\boldsymbol{X})$和$h_j(\boldsymbol{X})$称为约束函数．

因为$\min f(\boldsymbol{X})=-\max(-f(\boldsymbol{X}))$，所以，极大和极小本质上是相同的，而一个等式约束$h_j(\boldsymbol{X})=0$等价于两个不等式约束$h_j(\boldsymbol{X})\geqslant 0$与$-h_j(\boldsymbol{X})\geqslant 0$，因此上述数学规划问题可以写成统一的形式

$$\begin{aligned}(\text{MP})\qquad &\min f(\boldsymbol{X})\\ &g_i(\boldsymbol{X})\geqslant 0 \quad i=1,2,\cdots m\end{aligned}$$

或

$$(\mathrm{MP}) \quad \min f(\boldsymbol{X})$$

$$\boldsymbol{X} \in D$$

$$D = \{\boldsymbol{X} \mid g_i(\boldsymbol{X}) \geqslant 0 \quad i = 1,2,\cdots,m\}$$

根据变量 X 有无约束条件可分为约束规划问题和无约束规划问题．如果目标函数 $f(X)$ 与约束函数 $g_i(X)$ 都是线性函数，则相应的优化问题称为线性规划，否则称为非线性规划．

实际上除了上述两种类型的优化模型之外，还有其他的优化模型，如整数规划模型、多目标规划模型、动态规划模型．线性规划模型求解一般相对比较简单，可以采用单纯形法；非线性规划模型求解一般比较复杂，还有很多近似算法．同时也有很多优化的专用软件可以利用．

在充分理解问题的基础上，建立一个优化模型一般可遵循以下三个步骤：

(1)确定决策变量．对于一个决策问题，首先要明确待决策的内容或对象，然后设法将其变量化以确定决策变量．小型的线性规划问题可以只有较少的决策变量，大型问题则可能有上百个乃至成千上万个决策变量；有些问题的决策变量显而易见，有些则需要转化才能设出．

(2)确定目标函数．决策人面临着决策问题时，有一个进行方案决择的标准，即目标．在确定决策变量后，将决策目标表为决策变量的函数并根据实际问题求其最小或最大，即得目标函数．

(3)确定约束条件．决策问题的约束条件是指在决策过程中决策变量受到一定条件的限制，或达到一些平凡意义下最低限度的要求．它们的数学表现形式往往是决策变量的等式或不等式，线性规划中则体现为决策变量的线性等式或线性不等式．例如在生产经营的管理中，对资源(人力、物力、财力等)的使用常常是受到限制的，其资源使用量一般要受一定总量(上限)的限制或被要求满足最低使用量(下限)的要求．

1.3.5.1 线性规划模型

某公司面临一个是外包协作还是自行生产的问题．该公司生产甲、乙、丙三种产品，这三种产品要经过铸造、机加工和装配三个车间．甲、乙两种产品的铸件可以外包协作，亦可自行生产，但丙产品必须本厂铸造才能保证质量，有关情况见表 1.6．公司中可用的总工时为：铸造 8000 小时、机加工 12000 小时和装配 10000 小时．公司为了获得最大利润，甲、乙、丙三种产品各应该生产多少件？甲、乙两种产品的铸造应该多少由本公司铸造，多少外包协作？

表 1.6 公司产品生产安排及售价

		甲	乙	丙
2 时(小时/件)	铸造工时	5	10	7
	机加工工时	6	4	8
	装配工时	3	2	2
成本(百元/件)	自产铸件成本	3	5	4
	外协铸件成本	5	6	
	机加工成本	2	1	3
	装配成本	3	2	2
产品售价(百元/件)		23	18	16

(1)问题分析. 问题的目标就是利润最大,要做的决策是生产计划,即三种产品各生产多少,甲乙两种内铸多少,外协多少. 当然生产计划受到三个车间总工时的限制. 假设外协铸造不受能力限制.

(2)模型建立. 假设 x_1、x_2、x_3 分别表示甲、乙、丙三种产品的生产件数,而 x_1、x_2 又分为公司内部和外部铸造(x_{11},x_{12})和(x_{21},x_{22}),然后再由厂内机加工和装配. 从而有

$$x_1 = x_{11} + x_{12}, x_2 = x_{21} + x_{22}.$$

下面计算产品的利润.

产品甲自制部分的利润=23−(3+2+3)=15,共 $15x_{11}$,

产品甲外铸部分的利润=23−(5+2+3)=13,共 $13x_{12}$,

产品乙自制部分的利润=18−(5+1+2)=10,共 $10x_{21}$,

产品乙外铸部分的利润=18−(6+1+2)=9,共 $9x_{22}$,

产品丙的利润=16−(4+3+2)=7,共 $7x_3$.

目标函数:$\max(15x_{11}+13x_{12}+10x_{21}+9x_{22}+7x_3)$

约束条件:铸造 $5x_{11}+10x_{21}+7x_3 \leqslant 8000$,

机加工 $6x_1+4x_2+8x_3 \leqslant 12000$,

装配 $3x_1+2x_2+2x_3 \leqslant 10000$.

另外,生产数量是非负数,甲乙两种产品的数量还要满足上面的等式关系. 综合上述分析得到如下线性规划模型:

$$\max z = 15x_{11} + 13x_{12} + 10x_{21} + 9x_{22} + 7x_3.$$

满足

$$\begin{cases} 5x_{11} + 10x_{21} + 7x_3 \leqslant 8000, \\ 6x_1 + 4x_2 + 8x_3 \leqslant 12000, \\ 3x_1 + 2x_2 + 2x_3 \leqslant 10000, \\ x_1 = x_{11} + x_{12}, \\ x_2 = x_{21} + x_{22}, \\ x_1, x_2, x_3, x_{11}, x_{12}, x_{21}, x_{22} \geqslant 0. \end{cases}$$

1.3.5.2 非线性规划模型

某公司用两种原油(A 和 B)混合加工成两种汽油(甲和乙). 甲乙两种汽油含原油 A 的最低比例分别为 50%和 60%,每吨售价分别为 4800 元和 5600 元. 该公司现有原油 A 和 B 的库存量分别为 500 吨和 1000 吨,还可以从市场上买到不超过 1500 吨的原油 A. 原油 A 的市场价为:购买量不超过 500 吨时的单价为 10000 元(每吨);购买量超过 500 吨但不超过 1000 吨时,超过 500 吨的部分 8000 元(每吨);购买量超过 1000 吨时,超过 1000 吨的部分为 6000 元(每吨). 该公司该如何安排原油的采购与加工?

(1)问题分析. 安排原油采购和加工的目标是利润最大,题目中给出的是两种汽油的售价和原油 A 的采购价,利润为销售汽油的收入和购买原油 A 的支出之差. 这里的难点在于原油 A 的采购价与购买数量的关系比较复杂.

(2)模型建立. 设原油 A 的购买量为 x,根据题目中给出的数据,采购支出 $c(x)$可以表示为如下的分段函数(以下价格以千元(每吨)为单位):

$$c(x)=\begin{cases}10x & (0\leqslant x\leqslant 500),\\ 1000+8x & (500\leqslant x\leqslant 1000),\\ 3000+6x & (1000\leqslant x\leqslant 1500).\end{cases}$$

设原油 A 用于生产甲乙两种汽油的数量分别为 x_{11}和 x_{12},原油 B 用于生产甲乙两种汽油的数量分别为 x_{21}和 x_{22},则总的收入为 $4.8(x_{11}+x_{21})+5.6(x_{12}+x_{22})$,于是目标函数为

$$\max z=4.8(x_{11}+x_{21})+5.6(x_{12}+x_{22}).$$

约束条件包括两种汽油用的原油库存量的限制和原油 A 购买量的限制,以及两种汽油含原油 A 的比例限制,它们可以表示为

$$\begin{cases}x_{11}+x_{12}\leqslant 500+x,\\ x_{21}+x_{22}\leqslant 1000,\\ x\leqslant 1500,\\ \dfrac{x_{11}}{x_{11}+x_{21}}\geqslant 0.5,\\ \dfrac{x_{12}}{x_{12}+x_{22}}\geqslant 0.6,\\ x_{11},x_{12},x_{21},x_{22},x\geqslant 0.\end{cases}$$

上述模型显然是非线性规划模型.

又比如在实验数据处理或统计资料分析中,常常遇到这样的问题:如何利用有关变量的实验数据(资料)去确定这些变量间的函数关系. 例如,已知某物体的温度 φ 与时间 t 之间有如下形式的经验函数关系:

$$\varphi(t)=c_1+c_2t+\mathrm{e}^{c_3t}.$$

其中,c_1、c_2、c_3 是待定参数. 通过测试获得 n 组 φ 与 t 之间的实验数据$(t_i,\varphi_i)(i=1,2,\cdots,n)$,试确定参数 c_1、c_2、c_3,使理论曲线 $\varphi(t)=c_1+c_2t+\mathrm{e}^{c_3t}$尽可能地与 n 个测试点拟合.

考虑在最小二乘法的意义下确定最优参数 c_1、c_2、c_3. 为此,应该使理论曲线与实验曲线在测试点的偏差(误差)平方和取得最小值,即

$$\min\delta=\sum_{i=1}^{n}[\varphi_i-(c_1+c_2t_i+\mathrm{e}^{c_3t_i})]^2$$

这也是非线性规划问题.

上述两个模型里面只有一个目标函数. 事实上,还可以出现多个目标函数的情况.

1.3.5.3 多目标规划模型

某企业投资拟生产 A 和 B 两种产品,其投资费用分别为 2100 元(每吨)和 4800 元(每吨). 两种产品 A 和 B 的利润分别为 3600 元(每吨)和 6500 元(每吨). A 和 B 每月的最大生产能力

分别为 5 吨和 8 吨；市场对两种产品总量的需求每月不少于 9 吨．问该企业应该如何安排生产计划，才能既满足市场需求，又节约投资，而且使生产利润达到最大？

(1)问题分析．本例中实际上有两个要求，既要投资最小，而又要利润最大．

(2)模型建立．设 x_1 和 x_2 分别表示产品 A，B 每月的生产量(吨)；$f_1(x_1,x_2)$ 表示两种产品的总投资费用，$f_2(x_1,x_2)$ 表示两种产品获得总利润，目标函数为

$$\min f_1(x_1,x_2) = 2100x_1 + 4800x_2,$$

$$\max f_2(x_1,x_2) = 3600x_1 + 6500x_2.$$

约束条件为

$$\begin{cases} x_1 \leqslant 5, \\ x_2 \leqslant 8, \\ x_1 + x_2 \geqslant 9, \\ x_1, x_2 \geqslant 0. \end{cases}$$

上述例子尽管是比较简单的数学模型，但是这些模型的得来也很不容易，而且这些模型在当时的实际中也都发挥了重要作用，有些模型就是现在还在继续发挥作用．在工程实际中，还有很多种数学模型．一般而言，建立一个比较好的数学模型比较困难．在建立了数学模型之后，就要求解数学模型．由于数学模型的种类不一样，求解方法也千差万别．有些模型可以求得准确解，有些只能求得近似解．越复杂的问题一般来说也只能求得近似解．在下面章节的介绍中，将介绍一些可以说是比较通用的数值计算方法，这些方法在求解数学模型过程中也经常被用到．

习　题

1. 人带着猫、鸡和米过河，船除需要人划之外，至多能载猫、鸡和米三者之一，而当人不在场时，猫要吃鸡，鸡要吃米．试设计一个安全渡河方案，并使渡河次数尽量少．

2. 一家保姆服务公司专门向顾主提供保姆服务．根据估计，上一年的需求是：春季 6000 人·天，夏季 7500 人·天，秋季 5500 人·天，冬季 9000 人·天．公司新招聘的保姆必须经过 5 天的培训才能上岗，每个保姆每季度工作(新保姆包括培训)65 天．保姆从该公司而不是从顾主那里得到报酬，每人每月工资 800 元．春季开始时公司拥有 120 名保姆，在每个季度结束后，将有 15%的保姆自动离职．

(1)如果公司不允许解雇保姆，请你为公司制订下一年的招聘计划；哪些季度需求的增加不影响招聘计划？可以增加多少？

(2)如果公司在每个季度结束后允许解雇保姆，请为公司制订下一年的招聘计划．

3. 某钢管零售商从钢厂进货，将钢管按照顾客的要求进行切割后售出．从钢管厂进货时得到的原料钢管长度都是 1850mm．现有一客户需要 15 根 290mm、28 根 315mm、21 根 350mm 和 30 根 455mm 的钢管．为了简化生产过程，规定所使用的切割模式的种类不能超过 4 种，使用频率最高的一种切割模式按照一根原料钢管价值的 1/10 增加费用，使用频率次之的切割模式按照一根原料钢管价值的 2/10 增加费用，依次类推，且每种切割模式下的切割次

数不能太多(一根原料钢管最多生产 5 根产品). 此外,为了减少余料浪费,每种切割模式下的余料浪费不能超过 100mm. 为了使总费用最小,应如何下料?

4. 建立铅球投掷模型. 不考虑阻力,设铅球初速度为 v,出手高度为 h,出手角度为 α(与地面夹角),建立投掷距离与 v、h、α 的关系式,并在 v、h 一定的情况下求最佳出手角度.

5. 为了考评教师的教学质量,教学研究部门设计了一个教学评估表(表 1.7),对学生进行一次问卷调查,要求学生对 12 位教师的 15 门课程(其中 3 位教师有两门课)按一下内容打分,分值为 1 ~ 5 分(5 分最好,1 分最差). 其中,A 表示课程内容组织的合理性,B 表示主要问题展开的逻辑性,C 表示回答学生问题的有效性,D 表示课后交流的有助性,E 表示教科书的帮助性,F 考试评分的公正性,G 表示对教师的总体评价 . 收回问卷调查表后,得到了学生对 12 位教师、15 门课程各项评分的平均值 . 教学研究部门认为,所列各项具体内容 A ~ F 不一定对教师总体评价 G 有显著影响,并且各项内容之间也可能存在很强的相关性,他们希望得到一个总体评价与各项内容之间的模型,这个模型应该尽量简单有效,并且能给教师一些合理的建议,以提高总体评价.

表 1.7　教师及相关信息表

教师编号	课程编号	A	B	C	D	E	F	G
1	201	4.46	4.42	4.23	4.10	4.56	4.37	4.11
2	224	4.11	3.82	3.29	3.60	3.99	3.82	3.38
3	301	3.58	3.31	3.24	3.76	4.39	3.75	3.17
4	301	4.42	4.37	4.34	4.40	3.63	4.27	4.39
5	301	4.62	4.47	4.53	4.67	4.63	4.57	4.69
6	309	3.18	3.82	3.92	3.62	3.50	4.14	3.25
7	311	2.47	2.79	3.58	3.50	2.84	3.84	2.84
8	311	4.29	3.92	4.05	3.76	2.76	4.11	3.95
9	312	4.41	4.36	4.27	4.75	4.59	4.11	4.18
10	312	4.59	4.34	4.24	4.39	2.64	4.38	4.44
11	333	4.55	4.45	4.43	4.57	4.45	4.40	4.47
12	424	4.67	4.64	4.52	4.39	3.48	4.21	4.61
3	351	3.71	3.41	3.39	4.18	4.06	4.06	3.17
4	411	4.28	4.45	4.10	4.07	3.76	4.43	4.15
9	424	4.24	4.38	4.35	4.48	4.15	4.50	4.33

6. 高等教育学费标准探讨(2008 年高教社杯全国大学生数学建模竞赛试题). 高等教育事关高素质人才培养、国家创新能力增强、和谐社会建设的大局,因此受到党和政府及社会各方面的高度重视和广泛关注. 培养质量是高等教育的一个核心指标,不同的学科、专业在设定不同的培养目标后,其质量需要有相应的经费保障. 高等教育属于非义务教育,其经费在世界各国都由财政拨款、学校自筹、社会捐赠和学费收入等部分组成. 对适合接受高等教育的经济困难的学生,一般可通过贷款和学费减、免、补等方式获得资助,品学兼优者还能享受政府、学校、

企业等给予的奖学金. 学费问题涉及每一个大学生及其家庭，是一个敏感而又复杂的问题：过高的学费会使很多学生无力支付，过低的学费又使学校财力不足而无法保证质量. 学费问题近年来在各种媒体上引起了热烈的讨论. 请根据中国国情，收集如国家生均拨款、培养费用、家庭收入等相关数据，并据此通过数学建模的方法，就几类学校或专业的学费标准进行定量分析，并得出明确的、有说服力的结论. 数据的收集和分析是建模分析的基础和重要组成部分. 根据建模分析的结果，给有关部门写一份报告，提出具体建议.

2 数值计算方法概论

2.1 数值计算方法的研究对象和特点

数值计算方法，或者说数值方法，其传统的称呼为数值分析，又称为计算方法，它是伴随着计算机的出现和大规模计算的需求而发展起来的．数值方法主要研究各种数学模型及其算法，这些数学模型是为了解决各类应用领域特别是科学与工程计算领域的实际问题而提出的，因此，数值计算方法又称为科学计算(Scientific Computing)或计算科学工程(Computational Science and Engineering)．现在把科学计算看做是与理论及实验同等重要且必不可少的手段．随着科学技术的发展，计算机的性能和算法的效率都有了飞速的提高，要求解决的实际问题的规模也成倍扩大，其数学模型也日趋复杂．通常这些数学模型不能够准确求解，而只能借助于数值解法，然后在计算机上实现并做实际检验．这种数值模拟使人们能够研究复杂的系统和自然现象，而如果通过直接的实验来研究将是费时费钱或者非常危险的，有时甚至是不可能的．也正是数值模拟中从未有过的高层次的细节和现实性的探求，给计算机算法和系统结构中的重大突破提供了推动力．现在，随着计算能力的提高(包括硬件性能提高及各种高效算法的出现)，计算科学家和工程师们能过解决过去认为难以对付的问题，同时也期待解决一些超大规模的挑战性问题如基因测序、全球天气模拟等问题．

数值计算方法的主要任务是设计高效可靠的数值算法．对于同一个问题，不同的算法在计算性能上可能相差百万倍甚至更多．

例如当 $n=20$ 时，用 Cramer 法则解方程组，其运算次数(乘除法)需要 9.7×10^{20} 次，用每秒运算 1 亿次的计算机要用 30 多万年，而用 Gauss 消元法只需要乘除计算 2660 次，需要几秒钟而已．该例表明研究算法的重要性，同时它也表明只提高计算机的处理速度而不改进或选用好的算法也是不行的．人类的计算能力是计算工具的性能与计算方法效率的总和．因此，计算能力的提高依赖于双方的提高．由于算法研究所需要付出的代价要小得多，所以从某种意义上来看，研究和选择好的算法对提高计算速度比提高计算机的处理速度更重要．当然选择好算法的前提是要保证计算结果的可靠性，这就需要可靠的理论分析．不仅如此，在设计和选用算法时，还需要考虑一些其他因素．总体来讲，应该包括以下几个方面：

(1)可靠的理论分析，能任意逼近并达到精度要求．对近似算法要保证其收敛性和数值稳定性，还要进行误差分析．

(2)计算复杂性好，它包括时间复杂性和空间复杂性．在同一精度下，计算时间少的为时间复杂性好，而占用内存空间小的为空间复杂性好．

(3)数值试验，即通过数值试验证明一个算法是行之有效的．

总之，一个面向计算机，计算复杂性好，具有可靠的理论分析并通过数值试验检验过的算法就是一个好算法．

数值计算方法主要研究各种数学模型及其算法．算法是借助于计算机求解实现的一种大致思路，算法不等同于程序．这里所谈的数值计算方法是借助于数学和计算机解决问题的一

种思路，方法不等同于算法．只要理解清楚了数值求解的方法，写出算法是容易的．所以本书的侧重点是搞清楚数值计算方法本身的思想和原理．不过在本书中有些地方，算法就等同于方法．

为了进行高效地数值计算，已经研制了高效的强有力的数学软件如 Maple、Mathematica 和 Matlab，其中 Matlab 是工程计算界广泛流行的软件．不仅如此，还有针对特定问题的专用软件（如有优化中的 Lindo/Lingo 软件）．有时需要自己去编写程序．为了能有效掌握并应用这些程序或编写程序，必须首先学习科学计算中的一些基本概念、方法和理论，以及一些数值方法的主要思想和步骤．

2.2 数值计算方法的误差分析

在求解实际问题中，从建立模型，到将数学问题转化为数值问题，设计算法，并在计算机上进行实现，每一步都存在误差．

2.2.1 误差的来源

2.2.1.1 模型误差

为了用数值方法求解实际问题，首先要忽略一些次要因素，提出简化假设，在这些假设条件下，就可用简单的数学语言来描述实际问题和物理现象各种变量之间的关系，建立一个数学模型．显然这种数学模型是原问题的一个近似，它们之间的误差就称为模型误差．数值方法无法改进由模型误差产生的解的不准确性．因此数值方法中不讨论模型误差，通常都假设数学模型是合理的．

2.2.1.2 观测误差

在数学模型中往往还有一些根据观测得到的物理量，如温度、长度、电压等，通过观测的物理量要受到仪器精度的影响，因此这些参量中也包含误差，称为观测误差．这种误差会影响以这些数据为基础的任何计算解的精度．在数值计算方法中也不讨论观测误差．

2.2.1.3 截断误差

当数学模型不能得到精确解时，通常要采用数值计算方法求它的近似解，近似解和精确解之间的误差称为截断误差或方法误差．

例 2.1 函数 $f(x)$ 用泰勒多项式

$$P_n(x) = f(x_0) + f'(x_0)(x-x_0) + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n$$

来近似代替，则数值计算方法的截断误差为

$$R_n(x) = f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)^{n+1}.$$

其中，ξ 在 x_0 与 x 之间．

2.2.1.4 舍入误差

问题的数值计算方法确定后，具体需要用计算机求解，而由于机器字长有限，原始数据在

计算机上表示时会产生误差，计算过程又可能产生新的误差，这种误差称为舍入误差．例如用3.14159近似代替π产生的误差

$$R = \pi - 3.14159 = 0.0000026\cdots$$

就是舍入误差．

2.2.2 截断误差分析

截断误差是由一种近似的数值计算方法来求解数学问题而产生的误差．在例2.1中$R_n(x)\frac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)^{n+1}$是近似代替的余项，也就是用$P_n(x)$近似$f(x)$而产生的截断误差．由于$\xi$的值是未知的，这个截断误差的真正大小是不清楚的．因此，截断误差的余项$R_n(x)$具有理论上的意义．一般而言，有两种方法来大致估计截断误差．

(1)若存在正数$M>0$，满足对$\forall x\in[a,b]$，$|f^{(n+1)}(x)|\leqslant M$，则有

$$|R_n(x)|\leqslant\frac{M}{(n+1)!}|x-x_0|^{n+1}.$$

(2)若$n\to\infty$时，$R_n(x)\to 0$，则Taylor展开的级数收敛于$f(x)$．若记$h=x-x_0$，这时$R_n(x)$是一个与h^{n+1}同阶的无穷小量，记为$R_n(x)=O(h^{n+1})$(注意这里的“O”是大写的)．称用$P_n(x)$代替$f(x)$的截断误差是$O(h^{n+1})$，它表示用近似方法计算$f(x)$的精确度是$(n+1)$阶的．从而Taylor展开逼近可以表示成

$$f(x) = P_n(x) + O(h^{n+1}).$$

下面简单地补充说明关于记号小“o”和大“O”的用法，它们称为蓝道(Landau)记号，用来表示两个变量在极限过程中变化状态的相对关系，在数值计算方法中估计截断误差时经常用到．

(1)若$\lim\limits_{x\to x_0}f(x)=0$，即$f(x)$是$x\to x_0$时的无穷小量，记为$f(x)=o(1)$；

(2)若在x_0的某邻域内$|f(x)|\leqslant A$，即$f(x)$是$x\to x_0$时的一个有界量，记为$f(x)=O(1)$；

(3)若两个无穷小量$f(x)$和$g(x)$满足$\lim\limits_{x\to x_0}\frac{f(x)}{g(x)}=0$，即$f(x)$是比$g(x)$高阶的无穷小量，记为$f(x)=o(g(x))$；

(4)若两个无穷小量$f(x)$和$g(x)$满足$\lim\limits_{x\to x_0}\frac{f(x)}{g(x)}=c\neq 0$，即$f(x)$与$g(x)$是同阶的无穷小量，记为$f(x)=O(g(x))$.

另外，若$|f(x)|\leqslant c|g(x)$，$c>0$，也记为$f(x)=O(g(x))$.

按照定义有如下的运算关系式

(1) $o(1)\cdot O(f)=o(f)$，$O(1)\cdot O(f)=O(f)$；

(2) $O(f)\cdot O(g)=O(fg)$，$O(f)^n=O(f^n)$，n是正整数；

(3) $kO(f)=O(f)$，$O(kf)=O(f)$，k是常数；

(4) $O(f)+O(g)=O(f+g)$，$O(f)+O(f)=O(f)$；

(5) $O(f^n)+O(f^{n+1})=O(f^n)$，n是正整数．

例2.2 当$h\to 0$时，有如下Taylor展式

$$e^h = 1+h+\frac{h^2}{2!}+O(h^3), \sin(h) = h-\frac{h^3}{3!}+O(h^5),$$

试确定 $e^h+\sin(h)$ 和 $e^h\sin(h)$ 的截断误差.

解 (1) $e^h+\sin(h)=1+2h+\frac{h^2}{2!}-\frac{h^3}{3!}+O(h^3)+O(h^5)$.

注意到 $$-\frac{h^3}{3!}+O(h^3) = O(h^3), O(h^3)+O(h^5) = O(h^3),$$

从而 $$e^h+\sin(h) = 1+2h+\frac{h^2}{2!}+O(h^3).$$

上面的结果只是一个近似的估计. 如果将 e^h 展成

$$e^h = 1+h+\frac{h^2}{2!}+\frac{h^3}{3!}+O(h^4),$$

则有比较精确的阶的估计

$$e^h+\sin(h) = 1+2h+\frac{h^2}{2!}+O(h^4).$$

(2)类似于(1)可得

$$e^h\sin(h) = h+h^2+\frac{h^3}{3}+O(h^4).$$

若将 e^h 展成

$$e^h = 1+h+\frac{h^2}{2!}+\frac{h^3}{3!}+\frac{h^4}{4!}+O(h^5),$$

则有比较精确的阶的估计

$$e^h\sin(h) = h+h^2+\frac{h^3}{3}+O(h^5).$$

2.2.3 舍入误差分析

2.2.3.1 误差与有效数字

定义 2.1 设 x 为准确值,x^* 是近似值,称 $e^* = x-x^*$ 为近似值 x^* 的绝对误差,或简称为误差. 若 $x\neq 0$,称 $\frac{x-x^*}{x}$ 为近似值 x^* 的相对误差.

实际上准确值 x 往往是未知的,所以常把 $\frac{x-x^*}{x^*}$ 当做相对误差.

通常无法计算出准确值 x,也不能算出误差 e^* 的准确值,只能根据测量工具或计算情况估计出误差的绝对值不超过某一个正数 ε^*,所以给出如下定义.

定义 2.2 设 x 是准确值,x^* 是近似值,若存在正数 ε^* 满足

$$|x-x^*|\leqslant\varepsilon^*$$

称 ε^* 是近似值 x^* 的绝对误差限(或称为绝对误差界)(有时 ε^* 也记为 $\varepsilon(x^*)$),简称为误差限

(误差界). 相应地,称 $\varepsilon_r^* = \dfrac{\varepsilon^*}{|x|}$ 或 $\varepsilon_r^* = \dfrac{\varepsilon^*}{|x^*|}$ 为近似值的相对误差限.

2.2.3.2 有效数字

当准确值 x 有很多位时,经常按四舍五入的原则得到 x 的近似值 x^*. 不难验证,这样得到的近似值,其绝对误差限可以取为被保留的最后数位上的半个单位,例如

$$|\pi - 3.14| \leqslant 0.5 \times 10^{-2},\ |\pi - 3.142| \leqslant 0.5 \times 10^{-3},$$

$$|\pi - 3.1416| \leqslant 0.5 \times 10^{-4},\ |\pi - 3.14159| \leqslant 0.5 \times 10^{-5}.$$

由此可引入有效数字的概念.

定义 2.3 设 x 的近似值 x^* 有如下形式

$$x^* = \pm 10^m \times 0.a_1 a_2 \cdots a_n \cdots$$

其中 $a_i (i=1,2,\cdots)$ 是 0,1,2,…,9 中的一个数,$a_1 \neq 0$,m 为整数,若有

$$|x - x^*| \leqslant 0.5 \times 10^{m-n},$$

则称 x^* 为 x 的具有 n 位有效数字的近似值.

有效数字的误差限是末位数单位的一半,因其本身体现了误差界,从而有效数字的末位不能随意添零或减零.

例 2.3 按四舍五入原则写出下列各数具有 5 位有效数字的近似数:

$$187.9325, 0.03785551, 8.000033, 2.7182818.$$

解 按定义,上述各数具有 5 位有效数字的近似数分别是

$$187.93, 0.037856, 8.0000, 2.7183.$$

2.2.3.3 数值运算中的误差估计

假定一元函数 $f(x)$ 具有二阶连续导数,x^* 为 x 的近似值,以 $f(x^*)$ 代替 $f(x)$,其误差限为 $\varepsilon(f(x^*))$,由 Taylor 公式有

$$f(x) - f(x^*) = f'(x^*)(x - x^*) + \frac{f''(\xi)}{2}(x - x^*)^2,$$

ξ 介于 x 与 x^* 之间.

取绝对值有

$$|f(x) - f(x^*)| \leqslant |f'(x^*)| \varepsilon(x^*) + \frac{|f''(\xi)|}{2}\varepsilon^2(x^*),$$

忽略高阶项有

$$|f(x) - f(x^*)| \leqslant |f'(x^*)| \varepsilon(x^*).$$

因此,可取函数的误差限为

$$\varepsilon(f(x^*)) \approx |f'(x^*)| \varepsilon(x^*).$$

例 2.4 设 x^* 为 x 的近似值,证明 $\sqrt[n]{x^*}$ 的相对误差限大约是 x^* 的相对误差限的 $\dfrac{1}{n}$.

证明 令 $f(x)=x^{1/n}$，用 $\sqrt[n]{x^*}$ 代替 $\sqrt[n]{x}$ 而产生的误差限为

$$\varepsilon(f(x^*))\approx|f'(x^*)|\varepsilon(x^*)=\frac{1}{n}(x^*)^{\frac{1}{n}-1}\varepsilon(x^*)$$

从而相对误差限为

$$\varepsilon_r^*=\varepsilon(f(x^*))/(x^*)^{\frac{1}{n}}=\frac{1}{n}\frac{\varepsilon(x^*)}{x^*}$$

如果 f 是 n 元函数，例如计算 $A=f(x_1,x_2,\cdots,x_n)$. 自变量 $x_1,x_2,\cdots,x_n$ 的近似值分别为 $x_1^*,x_2^*,\cdots,x_n^*$，则 A 的近似值为 $A^*=f(x_1^*,x_2^*,\cdots,x_n^*)$，由 Taylor 展开式得到

$$\begin{aligned}A^*-A&=f(x_1^*,x_2^*,\cdots,x_n^*)-f(x_1,x_2,\cdots,x_n)\\&\approx\sum_{k=1}^{n}\left(\frac{\partial f(x_1^*,x_2^*,\cdots,x_n^*)}{\partial x_k}\right)(x_k^*-x_k),\end{aligned}$$

从而，误差限为

$$\varepsilon(A^*)\approx\sum_{k=1}^{n}\left|\frac{\partial f(x_1^*,x_2^*,\cdots,x_n^*)}{\partial x_k}\right|\varepsilon(x_k^*),$$

A^* 的相对误差限为

$$\varepsilon_r^*(A^*)=\frac{\varepsilon(A^*)}{|A^*|}.$$

特别地，可以把上述结果应用到二元函数 $f(x_1,x_2)=x_1\pm x_2,x_1x_2,\frac{x_1}{x_2}$ 上. 这样很容易得到两数和、差、积、商的绝对误差限估计，即

$$\varepsilon(x_1^*\pm x_2^*)=\varepsilon(x_1^*)+\varepsilon(x_2^*),$$

$$\varepsilon(x_1^*x_2^*)\approx|x_1^*|\varepsilon(x_2^*)+|x_2^*|\varepsilon(x_1^*),$$

$$\varepsilon(x_1^*/x_2^*)\approx\frac{|x_1^*|\varepsilon(x_2^*)+|x_2^*|\varepsilon(x_1^*)}{|x_2^*|^2}\quad(x_2^*\neq0).$$

例 2.5 设 $x_1\approx6.1025,x_2\approx80.115$ 均具有 5 位有效数字，试估计由这些数据计算 x_1x_2 的绝对误差限和相对误差限.

解 已知 $x_1^*=6.1025,x_2^*=80.115$，由题意得到

$$\varepsilon(x_1^*)=0.5\times10^{-4},\varepsilon(x_2^*)=0.5\times10^{-3}.$$

误差限为

$$\begin{aligned}\varepsilon(x_1^*x_2^*)&\approx|x_1^*|\varepsilon(x_2^*)+|x_2^*|\varepsilon(x_1^*)=6.1025\times0.5\times10^{-3}+80.115\times0.5\times10^{-4}\\&=7.057\times10^{-3},\end{aligned}$$

$$\begin{aligned}\varepsilon_r^*(x_1^*x_2^*)&=\frac{\varepsilon^*(x_1^*x_2^*)}{|x_1^*x_2^*|}\approx\frac{\varepsilon(x_2^*)}{|x_2^*|}+\frac{\varepsilon(x_1^*)}{|x_1^*|}=\frac{0.5\times10^{-3}}{80.115}+\frac{0.5\times10^{-4}}{6.1025}\\&=0.144344\times10^{-4}.\end{aligned}$$

例 2.6 测得某圆柱体底面半径 R 的近似值 $R^*=100\text{mm}$，高 h 的近似值 $h^*=50\text{mm}$. 若已知 $|R-R^*|\leqslant 0.5\text{mm}$，$|h-h^*|\leqslant 0.5\text{mm}$. 求体积 $V=\pi R^2 h$ 的绝对误差限和相对误差限.

解 $\dfrac{\partial V}{\partial R}=2\pi Rh$，$\dfrac{\partial V}{\partial h}=\pi R^2$，由公式可知

$$\varepsilon(V^*)\approx\left|\left(\frac{\partial V}{\partial R}\right)^*\right|\varepsilon(R^*)+\left|\left(\frac{\partial V}{\partial h}\right)^*\right|\varepsilon(h^*),$$

其中

$$\left(\frac{\partial V}{\partial R}\right)^*=2\pi R^*h^*,\left(\frac{\partial V}{\partial h}\right)^*=\pi(R^*)^2,$$

$$\varepsilon(R^*)=0.5,\varepsilon(h^*)=0.5.$$

从而

$$\varepsilon(V^*)\approx 10000\pi,V^*=\pi(R^*)^2h^*.$$

相对误差限为

$$\varepsilon_r^*(V^*)=\frac{\varepsilon(V^*)}{V^*}=0.02.$$

2.3 病态问题、数值稳定性和避免误差危害

2.3.1 病态问题与条件数

在计算函数值 $f(x)$ 时，若用 x^* 代替自变量 x，则其相对误差是 $\dfrac{x-x^*}{x}$，函数值的相对误差为 $\dfrac{f(x)-f(x^*)}{f(x)}$，两个相对误差之比为

$$\left|\frac{f(x)-f(x^*)}{f(x)}\right|\Big/\left|\frac{x-x^*}{x}\right|\approx\left|\frac{xf'(x)}{f(x)}\right|=C.$$

C 称为计算函数值问题的条件数. 对一般的问题，自变量的相对误差不会太大，如果条件数 C 很大，将引起函数值相对误差很大. 出现这种情况的问题称为病态问题. 一般情况下 $C\geqslant 10$，就认为是病态，C 越大，病态越严重.

例如 $f(x)=x^n$，则有 $C=n$，它表示相对误差可能放大 n 倍. 如 $n=10$，有 $f(1)=1$，$f(1.02)\approx 1.24$. 若取 $x=1$，$x^*=1.02$，则自变量相对误差为 2%，函数值相对误差为 24%，相对误差之比等于 12. 这时可以认为问题是病态的.

若 $n=0.5$，$f(x)=\sqrt{x}$，这时 $C=0.5$. 取 $x=1$，$x^*=1.02$，有 $f(1)=1$，$f(1.02)\approx 1.01$ 自变量的相对误差是 2%，函数值的相对误差是 1%，相对误差之比等于 0.5.

在其他的计算问题中也要分析是否病态的问题. 例如在解线性方程组时，如果输入数据（方程组系数和右端项）微小的误差引起解的巨大的误差，这就是病态方程组，后面将利用矩阵的条件数来大致分析. 由于初始数据总会有舍入误差，所以是否病态的问题应该引起足够重视. 一个问题是否病态，决定于问题的本身，与采用什么数值方法来求解并无关系.

2.3.2 数值计算方法的稳定性

误差的积累和传播是比较复杂的. 一种算法，若在一定条件下，其舍入误差在整个运算过

程中能够得到控制或者说舍入误差的增长不影响得到可靠的结果，则称该算法是稳定的，否则称其为不稳定的．下面看一个计算实例．

例 2.7 计算积分 $I_n=\int_0^1 \frac{x^n}{x+5}\,\mathrm{d}x$.

解 由于

$$I_n+5I_{n-1}=\int_0^1 \frac{x^n}{x+5}\mathrm{d}x+5\int_0^1 \frac{x^{n-1}}{x+5}\mathrm{d}x=\int_0^1 x^{n-1}\,\mathrm{d}x=\frac{1}{n},$$

从而

$$I_n=\frac{1}{n}-5I_{n-1}.$$

取 $I_0=\ln 6-\ln 5\approx 0.182$，逐步递推可得

$$I_1=0.090, I_2=0.050, I_3=0.083, I_4=-0.165,$$

$$I_5=1.025, I_6=-4.958, I_7=24.933, I_8=-124.540.$$

易见

$$\frac{1}{6(n+1)}=\int_0^1 \frac{x^n}{6}\mathrm{d}x\leqslant I_n\leqslant\int_0^1 \frac{x^n}{5}\mathrm{d}x=\frac{1}{5(n+1)}.$$

从而上述计算中，负的或大于 1 的结果都是错误的．上面的计算过程没有问题，是什么原因出现负的或大于 1 的计算结果呢？

对于准确值 I_n 而言，满足

$$I_0=\ln 6-\ln 5, I_n=\frac{1}{n}-5I_{n-1}.$$

对上述递推公式进行近似计算，近似计算值 I_n^* 满足

$$I_0^*=0.182, I_n^*=\frac{1}{n}-5I_{n-1}^*.$$

两式相减得到误差方程

$$I_n-I_n^*=(-5)(I_{n-1}-I_{n-1}^*),$$

递推得到

$$I_n-I_n^*=(-5)(I_{n-1}-I_{n-1}^*)=(-5)^2(I_{n-2}-I_{n-2}^*)=\cdots=(-5)^n(I_0-I_0^*).$$

从而得到递推计算的第 n 步的误差是初始误差的 $(-5)^n$ 倍，尽管初始误差 $|I_0-I_0^*|<0.5\times 10^{-3}$ 不大，但是随着 n 的增大，$|(-5)^n\times 0.5\times 10^{-3}|$ 就可以非常大．因此，这种算法是一种不稳定的算法．

如果换一种方式用递推公式

$$I_{n-1}=\frac{1}{5n}-\frac{1}{5}I_n\quad(n=8,7,\cdots,1),$$

取 $I_8\approx\frac{1}{2}\left(\frac{1}{6\times 9}+\frac{1}{5\times 9}\right)\approx 0.020$，逐步递推计算可得

$$I_7 = 0.021, I_6 = 0.024, I_5 = 0.028, I_4 = 0.034,$$

$$I_3 = 0.043, I_2 = 0.058, I_1 = 0.088, I_0 = 0.182.$$

这时

$$I_{n-1} - I_{n-1}^* = -\frac{1}{5}(I_n - I_n^*).$$

初始的误差在后面的计算中越来越小，得到了控制．这种算法是稳定的算法．

2.3.3 避免误差危害

误差的传播和积累在一些实际问题中是非常复杂的，不像例 2.7 可以得到一个误差传播的具体公式．在计算中，首先应分清问题是否病态和算法是否稳定，计算时还要尽量避免误差危害，防止有效数字的损失．通过研究和实际的计算，总结出下面一些基本的避免误差危害的原则．

2.3.3.1 避免相近的数相减

在数值计算中，两个相近的数相减时有效数字会损失．例如，计算

$$y = \sqrt{x+1} - \sqrt{x}.$$

其中，x 是比较大的数，例如 $x=1000$，取四位有效数字计算，有

$$y = \sqrt{1001} - \sqrt{1000} = 31.64 - 31.62 = 0.02.$$

在计算过程中，可以看到每个根号的计算都有四位有效数字，相减之后结果只有 1 位有效数字，相对误差变得很大，严重影响了结果的精确程度．事实上，可以采用如下等价计算公式：

$$y = \frac{1}{\sqrt{x+1} + \sqrt{x}}.$$

按此公式计算可得 $y=0.01581$，有四位有效数字．可见数学上等价的计算公式在实际计算中是不等价的．变换公式有很多种，例如：

$$\frac{1}{x} - \frac{1}{x+1} = \frac{1}{x(x+1)},$$

$$\ln(x+1) - \ln x = \ln\frac{x+1}{x},$$

$$\ln(x - \sqrt{x^2-1}) = -\ln(x + \sqrt{x^2-1}),$$

$$\sin(x+\varepsilon) - \sin x = 2\cos\left(x + \frac{\varepsilon}{2}\right)\sin\frac{\varepsilon}{2},$$

等等．当 x 比较大或 ε 比较小时，上述等式的右边的计算公式都要比左边的有效．

2.3.3.2 避免量级相差太大的两数相除

计算中大数除以小数或小数除以大数时，容易出现计算机溢出的情形．在这种情况下，有必要在量级上对这两个数进行一些处理．

2.3.3.3 避免大数和小数相加减

在数值计算中，有时会碰到数量级相差很大的两个数相加或相减．计算机做加法是要对阶的，即把这两个数都写成同一个阶数的表示形式，再对尾数相加减．

例如假设十进制 5 位数机器上做下面的加法：

$$12345+0.7.$$

计算机做加法时，要把这两数都写成尾数小于 1 的同阶的数，即

$$0.12345\times10^5+0.000007\times10^5.$$

但是计算机只能表示五位尾数，因此，第二个加数在计算机上就等于 0，这种情况称为“大数吃掉小数”

2.3.3.4 简化计算步骤

同样一个计算问题，如果能减少运算次数，不但可以节约计算机的计算时间，还能减少舍入误差．

例如计算多项式

$$P_n(x)=\alpha_n x^n+\alpha_{n-1}x^{n-1}+\cdots+\alpha_1 x+\alpha_0$$

的值，若直接计算 $\alpha_k x^k$ 再逐次相加，一共需要做

$$n+(n-1)+\cdots+2+1=\frac{n(n+1)}{2}$$

次乘法和 n 次加法．若采用秦九韶算法(Horner 算法)

$$\begin{cases}S_n=\alpha_n,\\ S_k=xS_{k+1}+\alpha_k \quad (k=n-1,n-2,\cdots,1,0),\\ P_n(x)=S_0.\end{cases}$$

只要 n 次乘法和 n 次加法就可计算出 $P_n(x)$的值．

习　题

1. 已知 e=2.7182818…，下列数作为 e 的近似值，试指出它们有几位有效数字？

$$x_1^*=2.71,x_2^*=2.7181,x_3^*=2.72.$$

2. 下列各数都是经过四舍五入得到的近似数，试指出它们有几位有效数字？

$$x_1^*=1.1021,x_2^*=0.031,x_3^*=385.6.$$

3. 设 x_1^*,x_2^*,x_3^* 按照第 2 题取值，求下列近似值的误差限：(1) $x_1^*+x_2^*+x_3^*$；(2) $x_1^*x_2^*x_3^*$．

4. 计算球体积要使相对误差限为 1%，问测量半径 R 允许的相对误差限是多少？

5. 已测得某场地长 l 的值 $l^*=110$m，宽 d 的值为 $d^*=80$m，已知 $|l-l^*|\leqslant0.2$m，$|d-d^*|\leqslant0.1$m，求面积 $A=ld$ 的绝对误差限和相对误差限．

6. 求方程 $x^2-56x+1=0$ 的两个根，使它至少具有 4 位有效数字（$\sqrt{783}\approx 27.982$）.

7. 当 N 充分大时，如何计算 $\int_N^{N+1}\frac{1}{1+x^2}\mathrm{d}x$?

8. 计算 $\alpha=(\sqrt{2}-1)^6$，取 $\sqrt{2}\approx 1.4$，分别采用下列等式计算：

(1) $\frac{1}{(\sqrt{2}+1)^6}$；　(2) $99-70\sqrt{2}$；　(3) $(3-2\sqrt{2})^3$；　(4) $\frac{1}{(3+2\sqrt{2})^3}$.

从算法设计原则上定性地判定哪一个将给出较好的近似值?

9. 序列 $\{x_n\}$ 满足

$$x_n = 10x_{n-1} - 1 \quad (n = 1,2,\cdots).$$

若取 $x_0=\sqrt{2}\approx 1.41$（三位有效数字），计算到 x_{100} 时误差有多大? 这个计算过程是否稳定?

10. 试导出计算积分

$$I_n = \int_0^1 \frac{x^n}{1+4x}\mathrm{d}x \quad (n = 1,2,\cdots)$$

的递推公式

$$I_n = \frac{1}{4}\left(\frac{1}{n} - I_{n-1}\right).$$

用此递推公式计算积分的近似值并分析计算误差．这个计算过程是否稳定?

3 插 值 法

3.1 引言

考虑表 3.1 中数据.

表 3.1 一维数据表

x_i	1.0	2.0	3.0	4.0	5.0	6.0
y_i	1.9	2.7	4.8	5.3	7.1	9.4

这些数据可能是某个函数在不同自变量处的函数值;也可能是实验得出的数据,例如设 x 表示温度,y 表示压力,即通过实验得出不同温度下的压力值;这些数据也可能表示一些自然现象中的数据,例如设 x_i 表示时间,y_i 表示不同时间点处的股票价格.

对于这些数据,可能希望划一条光滑曲线能通过这些点;也可能希望推测两个数据点之间的值或者预测数据表之外的某个 t 处的值. 如果上述数据是某个函数给出的信息,也可能希望计算出函数在某个点处的导数值或者函数在某个区间上的积分值.

当然有很多种方式可以来处理这些数据. 本章介绍用一种比较简单的函数来描述这些离散的数据,这种函数在图象上必须经过所有的离散点(x_j, y_i),也就是说函数在所有 x_i 处的值必须等于 y_i. 这种函数就是所谓的插值函数.

设函数 $y=f(x)$在区间$[a,b]$上有定义,且已知在点 $a\leqslant x_0<x_1<\cdots<x_n\leqslant b$ 上的值 y_0, $y_1,\cdots,y_n$,若存在一简单函数 $P(x)$,使

$$P(x_i)=y_i \quad (i=0,1,2,\cdots,n) \tag{3.1}$$

成立,就称 $P(x)$为 $f(x)$的插值函数,点 $x_0,x_1,\cdots,x_n$ 称为插值节点,包含插值节点的区间$[a,b]$称为插值区间,求插值函数 $P(x)$的方法称为插值法. 若 $P(x)$是次数不超过 n 的代数多项式,即

$$P(x)=a_0+a_1x+a_2x^2+\cdots+a_nx^n, \tag{3.2}$$

其中 a_i 为实数,就称 $P(x)$为插值多项式,相应的插值法称为多项式插值. 若 $P(x)$为分段的多项式,相应的插值法称为分段插值. 若 $P(x)$为三角多项式,相应的插值法称为三角插值.

对于多项式插值而言,由式(3.1)和式(3.2)可知,未知系数 $a_0,a_1,\cdots,a_n$ 满足

$$\begin{cases} a_0+a_1x_0+\cdots+a_nx_0^n=y_0, \\ a_0+a_1x_1+\cdots+a_nx_1^n=y_1, \\ \cdots\cdots \\ a_0+a_1x_n+\cdots+a_nx_n^n=y_n. \end{cases} \tag{3.3}$$

这是一个关于 $a_0,a_1,\cdots,a_n$ 的 $n+1$ 元线性代数方程组. 注意到方程组的系数行列式为范德

蒙行列式

$$V_n(x_0,x_1,\cdots,x_n)=\begin{vmatrix}1 & x_0 & x_0^2 & \cdots & x_0^n\\ 1 & x_1 & x_1^2 & \cdots & x_1^n\\ \vdots & \vdots & \vdots & & \vdots\\ 1 & x_n & x_n^2 & \cdots & x_n^n\end{vmatrix}=\prod_{i=1}^{n}\prod_{j=0}^{i-1}(x_i-x_j). \tag{3.4}$$

由于 $i\neq j$ 时，$x_i\neq x_j$，故所有因子 $x_i-x_j\neq 0$，于是 $V_n(x_0,x_1,\cdots,x_n)\neq 0$. 根据解线性方程组的克莱姆(Cramer)法则，方程组的解 a_i 存在唯一. 上述分析实质上给出了一个求多项式插值的一个方法，而且可以看出插值多项式是唯一的. 但是在节点 x_i 太多的情况下，范德蒙行列式的计算量太大，并不实用.

3.2 Lagrange 插值多项式

实际中一般采用先求插值基函数再求插值函数的方法. 下面讨论求通过 $n+1$ 个节点 $x_0<x_1<\cdots<x_n$ 的 n 次插值多项式 $L_n(x)$ 的办法，假定它满足条件

$$L_n(x_j)=y_j \quad (j=0,1,2,\cdots,n). \tag{3.5}$$

定义 3.1 若 n 次多项式 $l_j(x)(j=0,1,\cdots,n)$ 在 $n+1$ 个节点 $x_0<x_1<\cdots<x_n$ 上满足条件

$$l_j(x_k)=\begin{cases}1, & k=j\\ 0, & k\neq j\end{cases} \quad (j,k=0,1,\cdots,n). \tag{3.6}$$

就称这 $n+1$ 个 n 次多项式 $l_0(x),l_1(x),\cdots,l_n(x)$ 为节点 $x_0,x_1,\cdots,x_n$ 上的 n 次插值基函数. 容易得到

$$\begin{aligned}l_k(x)&=\frac{(x-x_0)\cdots(x-x_{k-1})(x-x_{k+1})\cdots(x-x_n)}{(x_k-x_0)\cdots(x_k-x_{k-1})(x_k-x_{k+1})\cdots(x_k-x_n)}\\ &=\prod_{\substack{j=0\\ j\neq k}}^{n}\frac{x-x_j}{x_k-x_j} \quad (k=0,1,\cdots,n).\end{aligned} \tag{3.7}$$

于是，满足条件式(3.5)的插值多项式 $L_n(x)$ 可表示为

$$L_n(x)=\sum_{k=0}^{n}y_k l_k(x). \tag{3.8}$$

由 $l_k(x)$ 的定义，知

$$L_n(x_j)=\sum_{k=0}^{n}y_k l_k(x_j)=y_j \quad (j=0,1,\cdots,n).$$

形如式(3.8)的插值多项式 $L_n(x)$ 称为拉格朗日(Lagrange)插值多项式. 显然由式(3.8)确定的多项式和由式(3.3)确定的多项式是一致的.

若 $n=1$，称为线性插值，$n=2$ 称为抛物型插值.

若在 $[a,b]$ 上用 $L_n(x)$ 近似 $f(x)$，则其截断误差为 $R_n(x)=f(x)-L_n(x)$，也称为插值多

项式的余项．关于插值余项估计有以下定理．

定理 3.1 设 $f^{(n)}(x)$在$[a,b]$上连续，$f^{(n+1)}(x)$在(a,b)内存在，节点 $a\leqslant x_0<\cdots<x_n\leqslant b$，$L_n(x)$是满足条件(3.5)的插值多项式，则对任何 $x\in[a,b]$，插值余项

$$R_n(x)=f(x)-L_n(x)=\frac{f^{(n+1)}(\xi)}{(n+1)!}\omega_{n+1}(x). \tag{3.9}$$

这里 $\xi\in(a,b)$且依赖于 $\omega_{n+1}(x)=(x-x_0)(x-x_1)\cdots(x-x_n)$．

证明 由给定条件知 $R_n(x)$在节点 $x_k(k=0,1,\cdots,n)$上为零，即

$$R_n(x_k)=0 \quad (k=0,1,\cdots,n).$$

于是

$$R_n(x)=K(x)(x-x_0)(x-x_1)\cdots(x-x_n)=K(x)\omega_{n+1}(x). \tag{3.10}$$

其中，$K(x)$是与 x 有关的待定函数．

现把 x 看成$[a,b]$上一个固定点，作函数

$$\varphi(t)=f(t)-L_n(t)-K(x)(t-x_0)(t-x_1)\cdots(t-x_n),$$

根据插值条件及余项定义，可知 $\varphi(t)$在点 $x_0,x_1,\cdots,x_n$ 及 x 均为零，故 $\varphi(t)$在$[a,b]$上有 $n+2$ 个零点．根据罗尔(Rolle)定理，$\varphi'(t)$在 $\varphi(t)$的两个零点间至少有一个零点，故 $\varphi'(t)$在$[a,b]$内至少有 $n+1$ 个零点．对 $\varphi'(t)$再应用罗尔定理，可知 $\varphi''(t)$在$[a,b]$内至少有 n 个零点．依此类推，$\varphi^{(n+1)}(t)$在(a,b)内至少有一个零点，记为 $\xi\in(a,b)$，使

$$\varphi^{(n+1)}(\xi)=f^{(n+1)}(\xi)-(n+1)!K(x)=0,$$

于是

$$K(x)=\frac{f^{(n+1)}(\xi)}{(n+1)!},\xi\in(a,b)\text{ 且依赖于 }x.$$

将它代入式(3.10)，就得到余项表达式——式(3.9)．

应当指出，余项表达式只有在 $f(x)$的高阶导数存在时才能应用．ξ 在(a,b)内的具体位置通常不可能给出，如果可以求出 $\max\limits_{a\leqslant x\leqslant b}|f^{(n+1)}(x)|=M_{n+1}$，那么插值多项式 $L_n(x)$逼近 $f(x)$的截断误差限是

$$|R_n(x)|\leqslant\frac{M_{n+1}}{(n+1)!}|\omega_{n+1}(x)|. \tag{3.11}$$

当 $n=1$ 时，线性插值余项为

$$R_1(x)=\frac{1}{2}f''(\xi)(x-x_0)(x-x_1)\quad\xi\in[x_0,x_1], \tag{3.12}$$

当 $n=2$ 时，抛物插值的余项为

$$R_2(x)=\frac{1}{6}f'''(\xi)(x-x_0)(x-x_1)(x-x_2)\quad\xi\in[x_0,x_2]. \tag{3.13}$$

例 3.1 若 sin0.32=0.314567，sin0.34=0.333487，sin0.36=0.352274，用线性插值及抛物插值计算 sin0.3367 的值并估计截断误差．

解 由题意取 $x_0=0.32, y_0=0.314567, x_1=0.34, y_1=0.333487, x_2=0.36, y_2=0.352274$. 用线性插值计算，取 $x_0=0.32$ 及 $x_1=0.34$，由式(3.7)和式(3.8)得

$$L_1(x)=y_0\frac{x-x_1}{x_0-x_1}+y_1\frac{x-x_0}{x_1-x_0}=y_0+\frac{y_1-y_0}{x_1-x_0}(x-x_0),$$

$$\sin 0.3367\approx L_1(0.3367)=y_0+\frac{y_1-y_0}{x_1-x_0}(0.3367-x_0)$$

$$=0.314567+\frac{0.01892}{0.02}\times 0.0167=0.330635.$$

其截断误差由式(3.12)得

$$|R_1(x)|\leqslant\frac{M_2}{2}|(x-x_0)(x-x_1)|,$$

其中

$$M_2=\max_{x_0\leqslant x\leqslant x_1}|f''(x)|.$$

注意到 $f(x)=\sin x, f''(x)=-\sin x$，取 $M_2=\max\limits_{x_0\leqslant x\leqslant x_1}|\sin x|=\sin x_1\leqslant 0.3335$，于是

$$|R_1(0.3367)|=|\sin 0.3367-L_1(0.3367)|$$

$$\leqslant\frac{1}{2}\times 0.3335\times 0.0167\times 0.0033\leqslant 0.92\times 10^{-5}$$

用抛物插值计算 $\sin 0.3367$ 时，由式(3.7)和式(3.8)得

$$L_2(x)=y_0\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}+y_1\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}+y_2\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)},$$

$$\sin 0.3367\approx L_2(0.3367)=0.330374.$$

这个结果与 6 位有效数字的正弦函数表完全一样，这说明查表时用二次插值精度已相当高了. 其截断误差限由式(3.13)得

$$|R_2(x)|\leqslant\frac{M_3}{6}|(x-x_0)(x-x_1)(x-x_2)|,$$

其中

$$M_3=\max_{x_0\leqslant x\leqslant x_2}|f'''(x)|=\cos x_0<0.828.$$

于是

$$|R_2(0.3367)|=|\sin 0.3367-L_2(0.3367)|$$

$$\leqslant\frac{1}{6}\times 0.828\times 0.0167\times 0.033\times 0.0233<0.178\times 10^{-6}.$$

本例中尽管 0.3367 处的正弦值是未知的，但是可以利用节点 0.32、0.34 和 0.36 处的正弦值通过做插值的手段来得到. 本例中第一种方法是利用前两个节点做线性插值函数 $L_1(x)$，并用 $L_1(0.3367)$ 来近似代替 $\sin 0.3367$. 第二种方法利用三个节点处的值做二次插值 $L_2(x)$，并用 $L_2(0.3367)$ 来近似代替 $\sin 0.3367$. 例子中也估计了两种方法的误差限. 另外 $\sin 0.3367$

$=0.3303741916$，所以近似计算值 $L_1(0.3367)$具有 3 位有效数字，$L_2(0.3367)$具有 6 位有效数字．

下面提出这样一个问题：能否通过本例中给出的数据按照同样的思路来计算 sin1.57 的近似值呢？即用 $L_1(1.57)$或 $L_2(1.57)$来近似代替 sin1.57.

经过计算可得：

$$L_1(1.57)=1.497067, L_2(1.57)=1.2414576249.$$

显然 sin1.57 的值不会超过 1．上面的近似计算方法对于前面的例子有效，而对于计算 sin1.57 则是不可取的计算方法．

3.3 牛顿插值

Lagrange 插值公式结构紧凑，表达式易于求得，理论分析非常方便，但是当插值节点发生变化时，Lagrange 插值基函数全部发生变化，整个插值公式也将发生变化，这在实际计算中很不方便．为了克服这一缺点，下面把插值多项式改写成

$$P_n(x)=a_0+a_1(x-x_0)+a_2(x-x_0)(x-x_1)+\cdots+a_n(x-x_0)\cdots(x-x_{n-1}),\tag{3.14}$$

其中 $a_0,a_1,\cdots,a_n$ 为待定系数，可由插值条件 $P_n(x_j)=f_j\quad(j=0,1,\cdots,n)$确定．

当 $x=x_0$ 时，$P_n(x_0)=a_0=f_0$.

当 $x=x_1$ 时，$P_n(x_1)=a_0+a_1(x-x_0)=f_1$，推得

$$a_1=\frac{f_1-f_0}{x_1-x_0}.$$

当 $x=x_2$ 时，$P_n(x_2)=a+a_1(x_2-x_0)+a_2(x_2-x_0)(x_2-x_1)=f_2$，推得

$$a_2=\frac{\dfrac{f_2-f_0}{x_2-x_0}-\dfrac{f_1-f_0}{x_1-x_0}}{x_2-x_1}.$$

依此递推可得到 $a_3,\cdots,a_n$．为写出系数 a_k 的一般表达式，先引进如下均差(也称为差商)定义．

定义 3.2 称

$$f[x_0,x_k]=\frac{f(x_k)-f(x_0)}{x_k-x_0}$$

为函数 $f(x)$关于点 x_0,x_k 的一阶均差．

$$f[x_0,x_1,x_k]=\frac{f[x_0,x_k]-f[x_0,x_1]}{x_k-x_1}$$

称为 $f(x)$的二阶均差．一般地称

$$f[x_0,x_1,\cdots,x_k]=\frac{f[x_0,\cdots,x_{k-2},x_k]-f[x_0,x_1,\cdots,x_{k-1}]}{x_k-x_{k-1}}\tag{3.15}$$

为 $f(x)$的 k 阶均差．均差有如下的基本性质：

性质 1 k 阶均差可表为函数值 $f(x_0),f(x_1),\cdots,f(x_k)$的线性组合，即

$$f[x_0,x_1,\cdots,x_k]=\sum_{j=0}^{k}\frac{f(x_j)}{(x_j-x_0)\cdots(x_j-x_{j-1})(x_j-x_{j+1})\cdots(x_j-x_k)}. \tag{3.16}$$

这个性质可用归纳法证明．这个性质也表明均差与节点的排列次序无关，称为均差的对称性．即

$$f[x_0,x_1,\cdots,x_k]=f[x_1,x_0,\cdots,x_k]=\cdots=f[x_1,\cdots,x_k,x_0].$$

性质 2 由性质 1 及式(3.15)可得

$$f[x_0,x_1,\cdots,x_k]=\frac{f[x_1,\cdots,x_k]-f[x_0,\cdots,x_{k-1}]}{x_k-x_0}. \tag{3.17}$$

性质 3 若 $f(x)$在$[a,b]$上存在 n 阶导数，且节点 $x_0,x_1,\cdots,x_n\in[a,b]$，则 n 阶均差与导数关系如下：

$$f[x_0,x_1,\cdots,x_n]=\frac{f^{(n)}(\xi)}{n!}\quad \xi\in[a,b], \tag{3.18}$$

这个公式可直接用罗尔定理证明．

均差计算可列均差表(表 3.2)．

表 3.2 均差表

x_k	$f(x_k)$	一阶均差	二阶均差	三阶均差	四阶均差
x_0	$f(x_0)$				
x_1	$f(x_1)$	$f[x_0,x_1]$			
x_2	$f(x_2)$	$f[x_1,x_2]$	$f[x_0,x_1,x_2]$		
x_3	$f(x_3)$	$f[x_2,x_3]$	$f[x_1,x_2,x_3]$	$f[x_0,x_1,x_2,x_3]$	
x_4	$f(x_4)$	$f[x_3,x_4]$	$f[x_2,x_3,x_4]$	$f[x_1,x_2,x_3,x_4]$	$f[x_0,x_1,x_2,x_3,x_4]$

根据均差定义，把 x 看成$[a,b]$上一点，可得

$$f(x)=f(x_0)+f[x,x_0](x-x_0),$$

$$f[x,x_0]=f[x_0,x_1]+f[x,x_0,x_1](x-x_1),$$

$$\cdots\cdots$$

$$f[x,x_0,\cdots,x_{n-1}]=f[x_0,x_1,\cdots,x_n]+f[x,x_0,\cdots,x_n](x-x_n).$$

把后一式代入前一式，得到

$$\begin{aligned}f(x)=&f(x_0)+f[x_0,x_1](x-x_0)+f[x_0,x_1,x_2](x-x_0)(x-x_1)+\cdots\\&+f[x_0,x_1,\cdots,x_n](x-x_0)(x-x_1)\cdots(x-x_{n-1})+f[x,x_0,\cdots,x_n]\omega_{n+1}(x)\\=&N_n(x)+R_n(x),\end{aligned}$$

其中

$$N_n(x)=f(x_0)+f[x_0,x_1](x-x_0)+f[x_0,x_1,x_2](x-x_0)(x-x_1)+\cdots$$
$$+f[x_0,x_1,\cdots,x_n](x-x_0)(x-x_1)\cdots(x-x_{n-1}),\tag{3.19}$$
$$R_n(x)=f(x)-N_n(x)=f[x,x_0,\cdots,x_n]\omega_{n+1}(x).\tag{3.20}$$

由式(3.19)确定的多项式 $N_n(x)$ 显然满足插值条件，且次数不超过 n，它就是形如式(3.14)的多项式，其系数为

$$a_k=f[x_0,x_1,\cdots,x_k].$$

$N_n(x)$ 称为牛顿(Newton)均差插值多项式．系数 a_k 就是各阶均差，它比拉格朗日插值计算量节省，且便于程序设计．

例 3.2 已知 $x=1,2,3,4,5$ 时，对应的函数值等于 1,4,7,8,6，求 4 次 Newton 插值多项式．

解 由给定的数据先构造均差表(表 3.3)．

表 3.3 均差表

x_k	$f(x_k)$	一阶均差	二阶均差	三阶均差	四阶均差
1	1				
2	4	3			
3	7	3	0		
4	8	1	−1	−1/3	
5	6	−2	−3/2	−1/6	1/24

从而

$$N_4(x)=1+3(x-1)+0(x-1)(x-2)-\frac{1}{3}(x-1)(x-2)(x-3)$$
$$+\frac{1}{24}(x-1)(x-2)(x-3)(x-4).$$

3.4 Hermite 插值

不少实际问题不但要求在节点上函数值相等，而且还要求对应的导数值也相等，甚至要求高阶导数也相等，满足这种要求的插值多项式就是埃尔米特(Hermite)插值多项式．下面只讨论函数值与导数值个数相等的情况．设在节点 $a\leqslant x_0<x_1<\cdots<x_n\leqslant b$ 上，$y_j=f(x_j)$，$m_j=f'(x_j)(j=0.1,2,\cdots,n)$，要求插值多项式 $H(x)$ 满足条件

$$H(x_j)=y_j,H'(x_j)=m_j(j=0,1,2,\cdots,n).\tag{3.21}$$

这里给出了 $2n+2$ 个条件，可唯一确定一个次数不超过 $2n+1$ 次的多项式 $H_{2n+1}(x)=H(x)$，其形式为

$$H_{2n+1}(x)=a_0+a_1x+\cdots+a_{2n+1}x^{2n+1},$$

如根据条件式(3.21)来确定 $2n+2$ 个系数 $a_0,a_1,\cdots,a_{2n+1}$ 显然非常复杂，因此，仍采用先求插值基函数方法．

先求插值基函数 $\alpha_j(x)$及 $\beta_j(x)(j=0,1,\cdots,n)$，共有 $2n+2$ 个，每一个基函数都是 $2n+1$ 次多项式，且满足条件

$$\begin{cases}\alpha_j(x_k)=\delta_{jk}=\begin{cases}0 & (i\neq k),\\ 1 & (j=k),\end{cases} \quad \alpha'_j(x_k)=0,\\ \beta_j(x_k)=0,\beta'_j(x_k)=\delta_{jk} \quad (j,k=0,1,\cdots,n).\end{cases} \tag{3.22}$$

于是满足条件式(3.22)的插值多项式 $H_{2n+1}(x)=H(x)$可写成用插值基函数表示的形式

$$H_{2n+1}(x)=\sum_{j=0}^{n}[y_j\alpha_j(x)+m_j\beta_j(x)]. \tag{3.23}$$

由条件式(3.22)，显然有 $H_{2n+1}(x_k)=y_k$，$H'_{2n+1}(x_k)=m_k$ $(k=0,1,\cdots,n)$可以求出 Hermite 插值的插值基函数 $\alpha_j(x)$及 $\beta_j(x)$为

$$\alpha_j(x)=\left[1-2(x-x_j)\sum_{\substack{k=0\\k\neq j}}^{n}\frac{1}{x_j-x_k}\right]l_j^2(x), \tag{3.24}$$

$$\beta_j(x)=(x-x_j)l_j^2(x), \tag{3.25}$$

其中 $l_j(x)$为对应节点 x_j 的 Lagrange 插值基函数.

还可证明满足条件式(3.21)的插值多项式是唯一的.

实际计算中对于只有两个节点 $x_0<x_1$ 的 Hermite 插值显得特别重要. 这时插值函数式(3.23)变成

$$\begin{aligned}H_3(x)=&y_0\left(1-2\frac{x-x_0}{x_0-x_1}\right)\left(\frac{x-x_1}{x_0-x_1}\right)^2+y_1\left(1-2\frac{x-x_1}{x_1-x_0}\right)\left(\frac{x-x_0}{x_1-x_0}\right)^2\\&+m_0(x-x_0)\left(\frac{x-x_1}{x_0-x_1}\right)^2+m_1(x-x_1)\left(\frac{x-x_0}{x_1-x_0}\right)^2.\end{aligned} \tag{3.26}$$

仿照 Lagrange 插值余项的证明方法可以证明，若 $f(x)$在(a,b)内存在 $2n+2$ 阶导数，则其插值余项

$$R(x)=f(x)-H_{2n+1}(x)=\frac{f^{(2n+2)}(\xi)}{(2n+2)!}\omega_{n+1}^2(x), \tag{3.27}$$

其中 $\xi\in(a,b)$且与 x 有关.

例 3.3 求满足 $P(x_j)=f(x_j)$ $(j=0,1,2)$及 $P'(x_1)=f'(x_1)$的插值多项式及其余项表达式.

解 由给定条件，可确定次数不超过 3 的插值多项式. 由于此多项式通过点$(x_0,f(x_0))$、$(x_1,f(x_1))$及$(x_2,f(x_2))$，故其形式为

$$\begin{aligned}P(x)=&f(x_0)+f[x_0,x_1](x-x_0)+f[x_0,x_1,x_2](x-x_0)(x-x_1)\\&+A(x-x_0)(x-x_1)(x-x_2),\end{aligned}$$

其中 A 为待定常数，可由条件 $P'(x_1)=f'(x_1)$确定. 通过计算可得

$$A = \frac{f'(x_1) - f[x_0, x_1] - (x_1 - x_0) f[x_0, x_1, x_2]}{(x_1 - x_0)(x_1 - x_2)}.$$

为了求出余项 $R(x) = f(x) - P(x)$ 的表达式，可设

$$R(x) = f(x) - P(x) = k(x)(x - x_0)(x - x_1)^2(x - x_2),$$

其中 $k(x)$ 为待定函数．构造

$$\varphi(t) = f(t) - P(t) - k(x)(t - x_0)(t - x_1)^2(t - x_2)$$

显然 $\varphi(x_j) = 0 (j = 0, 1, 2)$，且 $\varphi'(x_1) = 0$，$\varphi(x) = 0$，故 $\varphi(t)$ 在 (a, b) 内有 5 个零点（重根算两个）．反复应用罗尔定理，得 $\varphi^{(4)}(t)$ 在 (a, b) 内至少有一个零点 ξ，故

$$\varphi^{(4)}(\xi) = f^{(4)}(\xi) - 4! k(x) = 0,$$

于是

$$k(x) = f^{(4)}(\xi)/4!.$$

余项表达式为

$$R(x) = f^{(4)}(\xi)(x - x_0)(x - x_1)^2(x - x_2)/4!, \tag{3.28}$$

式中 ξ 位于 x_0, x_1, x_2 和 x 所界定的范围内．

3.5 分段线性插值

上面根据区间 $[a, b]$ 上给出的节点做插值多项式 $L_n(x)$ 近似 $f(x)$，一般总认为 $L_n(x)$ 的次数 n 越高，逼近 $f(x)$ 的精度越好，但实际上并非如此．这是因为对任意的插值节点，当 $n \to \infty$ 时，$L_n(x)$ 不一定收敛到 $f(x)$．20 世纪初龙格(Runge)就给出了一个等距节点插值多项式 $L_n(x)$ 不收敛到 $f(x)$ 的例子．高次插值多项式的这些缺陷，促使人们寻求简单的低次分段多项式插值．值得一提的是，分段插值也正是有限元方法的基础．下面只介绍最简单的分段线性插值．

所谓分段线性插值，就是通过插值点用折线段连接起来逼近 $f(x)$．设已知节点 $a \leqslant x_0 < x_1 < \cdots < x_n \leqslant b$ 上的函数值 $f_0, f_1, \cdots, f_n$，记 $h_k = x_{k+1} - x_k$，$h = \max\limits_k h_k$．求一折线函数 $I_h(x)$ 满足：

(1) $I_h(x)$ 在闭区间 $[a, b]$ 上连续；

(2) $I_h(x_k) = f_k (k = 0, 1, 2, \cdots, n)$；

(3) $I_h(x)$ 在每个区间 $[x_k, x_{k+1}]$ 上是线性函数．

则称 $I_h(x)$ 为分段线性插值函数．

由定义可知 $I_h(x)$ 在每个小区间 $[x_k, x_{k+1}]$ 上可表示为

$$I_h(x) = \frac{x - x_{k+1}}{x_k - x_{k+1}} f_k + \frac{x - x_k}{x_{k+1} - x_k} f_{k+1} \quad (x_k \leqslant x \leqslant x_{k+1}). \tag{3.29}$$

同理，还可以做出分段二次插值、分段 Hermite 插值．

3.6 样条插值

分段插值公式简单，运算量节省，稳定性好，收敛性有保证，且只要区间长度足够小，分段低次插值总满足精度要求．但分段插值的一个明显缺点是插值函数在节点处的导数值不连续，即插值不光滑．既要使用分段低次插值，同时使插值函数具有一定的光滑性，解决该问题的办法是使用样条插值，样条插值其实是一种改进的分段插值．下面介绍最常用的三次样条插值．

定义 3.3 若函数在区间$[a,b]$上给定节点$a=x_0<x_1<\cdots<x_n=b$及其函数值y_j，若函数$S(x)$满足：

(1)$S(x_j)=y_j(j=0,1,2\cdots,n)$；

(2)$S(x)$在每个小区间$[x_j,x_{j+1}](j=0,1,2,\cdots,n-1)$上是三次多项式；

(3)$S(x)$在区间$[a,b]$上有连续的二阶导数.

则称$S(x)$为$[a,b]$上的三次样条插值函数．

从定义可知，要求出$S(x)$，在每个小区间$[x_j,x_{j+1}]$上要确定 4 个待定系数，共有n个小区间，故应确定$4n$个参数．根据$S(x)$在$[a,b]$上二阶导数连续，在节点$x_j(j=1,2,\cdots,n-1)$处应满足连续性条件

$$S(x_j-0)=S(x_j+0),S'(x_j-0)=S'(x_j+0),S''(x_j-0)=S''(x_j+0),$$

共有$3n-3$个条件，再加上$S(x)$满足$n+1$个插值条件，共有$4n-2$个条件，因此还需要 2 个条件才能确定$S(x)$．通常可在区间$[a,b]$端点$a=x_0$、$b=x_n$上各加一个条件(称为边界条件)，可根据实际问题的要求给定．常见的有以下三种：

(1)已知两端的一阶导数值，即

$$S'(x_0)=f'_0,S'(x_n)=f'_n.$$

(2)已知两端的二阶导数值，即

$$S''(x_0)=f''_0,S''(x_n)=f''_n.$$

(3)当$f(x)$是以x_n-x_0为周期的周期函数时，则要求$S(x)$也是周期函数．这时边界条件应满足

$$S(x_0+0)=S(x_n-0),S'(x_0+0)=S'(x_n-0),S''(x_0+0)=S''(x_n-0),$$

而此时$y_0=y_n$．这样确定的样条函数$S(x)$称为周期样条函数．

利用插值条件、连续性条件及边界条件，就可以求出三次样条函数的表达式．

习　题

1. 依据表 3.4 建立不超过三次的 Lagrange 插值多项式和 Newton 插值多项式．

表 3.4　函数值表

x	0	1	2	4
$f(x)$	1	9	23	3

2. 设给定函数 $f(x)=\sin x$ 的正弦函数值表见表 3.5.

表 3.5　正弦函数值表

x	1.0	1.5	2.0
$\sin x$	0.8415	0.9775	0.9093

试用二次插值多项式计算 sin1.8 的近似值.

3. 已知多项式 $p(x)=x^4-x^3+x^2-x+1$ 通过表 3.6 中各点.

表 3.6　一维数据表

x	-2	-1	0	1	2	3
$p(x)$	31	5	1	1	11	61

试构造一多项式 $q(x)$通过表 3.7 中各点.

表 3.7　一维数据表

x	-2	-1	0	1	2	3
$q(x)$	31	5	1	1	11	1

4. 已知单调连续函数 $y=f(x)$的函数值见表 3.8.

表 3.8　一维函数值表

x	0	1	2
$f(x)$	8	-7.5	-18

求函数 $f(x)$在区间$[0,2]$之间的零点的近似值.

5. 设 $f(x)=x^7+x^4+1$,求 $f[2^0,2^1,\cdots,2^7]$及 $f[2^0,2^1,\cdots,2^8]$.

6. 依据表 3.9,构造次数不超过三次的插值多项式.

表 3.9　函数值及导数值

x	0	1
$f(x)$	1	0
$f'(x)$	0	-1

7. 求一个次数不高于四次的多项式 $P(x)$,使它满足 $P(0)=0,P'(0)=0,P(1)=1,P'(1)=1,P(2)=1$.

4 曲线拟合

4.1 引言

例 4.1 油气田和油气井的产量并不是恒定不变的，而是随着油气开采进程或开发措施的实施过程不断发生变化的．油气产量在一定时期内可能表现为上升趋势，而在另外的时期内则可能趋于稳定，但在油气田开发的大部分时间内，却以不断递减的趋势发展变化着．油气田开发递减期的长短主要受油气田地质条件和当时的经济技术条件的影响，大多数油气田都带有一个长长的产量递减期．一般情况下，油气田的产量递减期都在 10～30 年以上，递减期可以采出地质储量的 40%～50%左右．

递减期由于油气产量的不断减小，油气田的开发效益不断下滑．为了提高油气田开发的经济效益，一般情况下都要根据油气产量的递减规律，制订出相应的减缓产量递减的措施，因此，递减期的矿场工作量特别大，包括各种增产和增注措施的实施．待所有旨在提高油气开采效益的措施全部实施完毕以后，油气生产仍不能带来经济效益，油气开发过程将被终止，油气田最终被废弃．因此，研究产量递减规律对做好油气田的动态预测和油气生产规划，意义重大．同时，只有了解油气田的产量递减规律之后，才能有的放矢地采取防止产量递减的有效措施，以提高油气采收率．

根据大量的统计发现，油气产量的递减率(D)和产量(q)之间的统计关系满足

$$D = Kq^n.$$

若给定历史生产数据(t_i, q_i)(这时可以计算出和 t_i 相对应的 D_i)$(i=1,2,\cdots,M)$，如何确定上式中的参数 K、n?

除此之外，还有很多数学关系式可以用来描述产量随时间递减的趋势，如

$$q = a + \frac{b}{t},$$

$$q = a + \frac{b}{t^2},$$

$$q = a - b\ln t.$$

只要通过历史生产数据(t_i, q_i)能把上述公式中的常数 a、b 确定出来，就可以用来研究产量的变化规律．

例 4.2 在化学工程中经常会遇到计算高温状态下蒸气压和温度的问题，但是考虑到测量设备等的限制，希望利用低温状态下的蒸气压等有关数据进行外推．表 4.1 给出了氨蒸气的一组温度和蒸气压数据，那么能否从所列的数据中计算出 75℃氨蒸气压?

表 4.1 温度与蒸气压数据表

温度	20	25	30	35	40	45	50	55	60
蒸气压	805	985	1170	1365	1570	1790	2030	2300	2610

上述两个引例都有一个共同点，那就是从给定的数据表中挖掘出一些信息．事实上，在工程实践和科学实验中，这种问题并不少见．处理这类问题的一个重要方法就是采用所谓的曲线拟合．这种方法在第 1 章所述的人口增长预测模型中已经用到了．

4.2 曲线拟合的最小二乘法

在科学实验的统计方法研究中，往往要从一组实验数据$(x_i,y_i)(i=1,2,\cdots,m)$中，寻找自变量 x 与因变量 y 之间的函数关系 $y=f(x)$．寻求函数关系可以采用前面的插值方法，但是对于节点个数多的时候容易出现数值振荡(龙格现象)；如果采用样条插值，计算量又太大．值得注意的是，这些实验数据或者是观测数据本身往往就有误差．正是由于实验或观测数据往往不准确，因此完全没有必要要求待求的函数关系式 $y=f(x)$经过所有点(x_i,y_i)，而只要求在给定点 x_i 上的误差(也称为残差)$\delta_i=f(x_i)-y_i(i=0,1,2,\cdots,m)$按某种标准最小，这就是所谓的曲线拟合法．这里所说的标准，通常采用以下三种衡量准则：

(1)误差的最大绝对值最小

$$\max_i\{|\delta_i|\}=\min.$$

(2)误差的绝对值之和最小

$$\sum_i|\delta_i|=\min.$$

(3)误差的平方和最小

$$\sum_i\delta_i^2=\min.$$

(1)和(2)中含有绝对值，不利于实际计算，一般按照(3)来确定参数，称为曲线拟合(数据拟合)的最小二乘法．

曲线拟合的一般提法是：对给定的一组数据(x_i,y_i)，$(i=0,1,2,\cdots,m)$要求在函数类 $\varphi=\mathrm{span}\{\varphi_0,\varphi_1,\cdots,\varphi_n\}$中找一个函数 $y=S^*(x)$，使误差平方和

$$\|\boldsymbol{\delta}\|_2^2=\sum_{i=0}^m\delta_i^2=\sum_{i=0}^m\omega_i[S^*(x_i)-y_i]^2=\min_{S(x)\in\varphi}\sum_{i=0}^m\omega_i[S(x_i)-y_i]^2,\tag{4.1}$$

其中，ω_i 是节点 x_i 处的权重，

$$\delta_i=S^*(x_i)-y_i,$$

$$\boldsymbol{\delta}=(\delta_0,\delta_1,\cdots,\delta_m)^{\mathrm{T}},$$

$$S(x)=a_0\varphi_0(x)+a_1\varphi_1(x)+\cdots+a_n\varphi_n(x)\quad(n<m),$$

$$S^*(x)=a_0^*\varphi_0(x)+a_1^*\varphi_1(x)+\cdots+a_n^*\varphi_n(x)\quad(n<m),$$

这就是一般的曲线拟合的最小二乘法．

在上述 $S(x)$的表达式中，$\varphi_0(x),\varphi_1(x),\cdots,\varphi_n(x)$是已知函数且线性无关．用最小二乘法求拟合曲线的问题，就是在形如 $S(x)$的表达式中找 $S^*(x)$，使式(4.1)中的 $\sum_{i=0}^m\omega_i[S(x_i)-y_i]^2$

达到最小值，因此求 $S^*(x)$ 的问题等价于求多元函数

$$I(a_0,a_1,\cdots,a_n)=\sum_{i=0}^{m}\omega_i\Big[\sum_{j=0}^{n}a_j\varphi_j(x_i)-y_i\Big]^2$$

的极小值问题．由求多元函数极值的必要条件，有

$$\frac{\partial I}{\partial a_k}=2\sum_{i=0}^{m}\omega_i\Big[\sum_{j=0}^{n}a_j\varphi_j(x_i)-y_i\Big]\varphi_k(x_i)=0\quad(k=0,1,\cdots,n).\tag{4.2}$$

若记

$$(\varphi_j,\varphi_k)=\sum_{i=0}^{m}\omega_i\varphi_j(x_i)\varphi_k(x_i),\tag{4.3}$$

$$(y,\varphi_k)=\sum_{i=0}^{m}\omega_i y_i\varphi_k(x_i)\equiv d_k\quad(k=0,1,\cdots,n).\tag{4.4}$$

则式(4.2)可改写为

$$\sum_{j=0}^{n}(\varphi_j,\varphi_k)a_j=d_k\quad(k=0,1,\cdots,n),\tag{4.5}$$

该方程称为法方程．法方程实际上是关于 $a_0,a_1,\cdots,a_n$ 一个线性代数方程组．写成矩阵形式为

$$\boldsymbol{G}a=d,$$

其中

$$\boldsymbol{a}=(a_0,a_1,a_2,\cdots,a_n)^{\mathrm{T}},$$

$$\boldsymbol{d}=(d_0,d_1,d_2,\cdots,d_n)^{\mathrm{T}},$$

$$\boldsymbol{G}=\begin{pmatrix}(\varphi_0,\varphi_0) & (\varphi_1,\varphi_0) & \cdots & (\varphi_n,\varphi_0)\\ (\varphi_0,\varphi_1) & (\varphi_1,\varphi_1) & \cdots & (\varphi_n,\varphi_1)\\ \vdots & \vdots & & \vdots\\ (\varphi_0,\varphi_n) & (\varphi_1,\varphi_n) & \cdots & (\varphi_n,\varphi_n)\end{pmatrix}.$$

显然系数矩阵是对称阵．在系数矩阵可逆的情况下，该方程组有唯一解，而且还可以证明它的解就是要求的 $a_0^*,a_1^*,\cdots,a_n^*$．由于该法方程是线性代数方程组，所以也称式(4.1)为线性最小二乘曲线拟合．事实上，这时 $S(x)=a_0\varphi_0(x)+a_1\varphi_1(x)+\cdots+a_n\varphi_n(x)$ 对 $a_0,a_1,\cdots,a_n$ 是线性的．

用最小二乘法求拟合曲线时，首先要确定 $S(x)$ 的表达式．这已经不单纯是数学问题了，还与所研究的实际问题有关系．关于确定拟合曲线的问题，一般可从以下几个方面入手：

(1)利用已知的结论确定拟合曲线的形式，例如由胡克定律可以知道在一定条件下，弹性体的应变与应力呈线性关系；

(2)从分析实验数据入手，通过描点作图的方式大致判断曲线的增减性、凹凸性等，然后选择合适的曲线进行拟合．

一般而言，对拟合得到的曲线还要进行实例验证，验证拟合曲线的可行性．有时，某些问

题的拟合曲线的选取不是唯一的，这时可以将得到的多种拟合曲线进行分析、比较或实例验证，综合判断后选择一种合适的曲线进行预测推断．

例 4.3 已知一组实验数据(表 4.2)，求它的拟合曲线．

表 4.2 实验数据表

x_i	1	2	3	4	5
y_i	4	4.5	6	8	8.5
ω_i	2	1	3	1	1

解 将所给数据在坐标纸上标出，可以看出，各点在一条直线附近，故选择线性函数作拟合曲线，即令 $S(x)=a_0+a_1x$，这里 $m=4,n=1,\varphi_0(x)=1,\varphi_1(x)=x$，故

$$(\varphi_0,\varphi_0)=\sum_{i=0}^{4}\omega_i=8,$$

$$(\varphi_0,\varphi_1)=(\varphi_1,\varphi_0)=\sum_{i=0}^{4}\omega_i x_i=22,$$

$$(\varphi_1,\varphi_1)=\sum_{i=0}^{4}\omega_i x_i^2=74,$$

$$d_0=(\varphi_0,y)=\sum_{i=0}^{4}\omega_i y_i=47,$$

$$d_1=(\varphi_1,y)=\sum_{i=0}^{4}\omega_i x_i y_i=145.5.$$

由式(4.5)得到

$$\begin{cases}8a_0+22a_1=47,\\22a_0+74a_1=145.5.\end{cases}$$

解得 $a_0=2.77,a_1=1.13$．于是所求拟合曲线为

$$S^*(x)=2.77+1.13x$$

例 4.4 设数据$(x_i,y_i)(i=0,1,2,3,4)$由表 4.3 给出，拟合曲线为 $y=ae^{bx}$，求参数 a、b.

表 4.3 实验数据表

i	0	1	2	3	4
x_i	1.00	1.25	1.50	1.75	2.00
y_i	5.10	5.79	6.53	7.45	8.46

解 显然，拟合曲线对参数而言不是线性形式．这时首先将拟合曲线取对数得到

$$\ln y=\ln a+bx.$$

令$\bar{y}=\ln y,A=\ln a$，这时拟合曲线变成$\bar{y}=A+bx$，这时 $\varphi_0(x)=1,\varphi_1(x)=x$，为了计算方便，将数据表变形得到表 4.4.

表 4.4　变形后的数据表

i	0	1	2	3	4
x_i	1.00	1.25	1.50	1.75	2.00
y_i	5.10	5.79	6.53	7.45	8.46
$\overline{y}_i$	1.629	1.756	1.876	2.008	2.135

数据表中没有给出权重 ω_i，约定取 $\omega_i=1$. 由式(4.3)和式(4.4)可得

$$(\varphi_0,\varphi_0)=5,$$

$$(\varphi_0,\varphi_1)=(\varphi_1,\varphi_0)=\sum_{i=0}^{4}x_i=7.5,$$

$$(\varphi_1,\varphi_1)=\sum_{i=0}^{4}x_i^2=11.875,$$

$$(\varphi_0,\overline{y})=\sum_{i=0}^{4}\overline{y}_i=9.404,$$

$$(\varphi_1,\overline{y})=\sum_{i=0}^{4}x_i\overline{y}_i=14.422.$$

法方程为

$$\begin{cases}5A+7.50b=9.404,\\7.50A+11.875b=14.422.\end{cases}$$

解得 $A=1.122$，$b=0.505$，$a=\mathrm{e}^A=3.071$，从而最小二乘曲线为

$$y=3.071\mathrm{e}^{0.505x}.$$

现在很多计算机配有自动选择数学模型的程序．程序中因变量与自变量变换的函数类型较多，通过计算比较误差找到拟合的较好的曲线，最后输出曲线图形及数学表达式．

对于非线性的拟合曲线，除了 $y=a\mathrm{e}^{bx}$ 外，经常用到的还有下面几种形式：

(1) $y=ax^b$　$(x>0,a>0)$；

(2) $y=a\mathrm{e}^{\frac{b}{x}}$　$(a>0)$；

(3) $y=\dfrac{1}{a+b\mathrm{e}^{-x}}$；

(4) $y=ax^b\mathrm{e}^{-\alpha}$　$(a>0)$.

有时，下面两种线性情况也是常用的拟合曲线：

(1) $\dfrac{1}{y}=a+\dfrac{b}{x}$；

(2) $y=a+b\ln x$.

如果是上面这些拟合曲线，该如何进行曲线拟合？特别地，如果拟合曲线是 $y=a\mathrm{e}^{bx}+c$（其中 a,b,c 为拟合参数），该如何进行拟合？

例 4.5 求超定方程组

$$\begin{cases} 2x_1 + 4x_2 = 11 \\ 3x_1 - 5x_2 = 3 \\ x_1 + 2x_2 = 6 \\ 2x_1 + x_2 = 7 \end{cases}.$$

的最小二乘解，并求误差的平方和．

解 显然方程组的第一个和第三个方程是不相容的，方程组肯定无解．方程组的最小二乘解就是在一个满足在误差的平方和的意义下是最小的近似解．令

$$I = (2x_1 + 4x_2 - 11)^2 + (3x_1 - 5x_2 - 3)^2 + (x_1 + 2x_2 - 6)^2 + (2x_1 + x_2 - 7)^2.$$

下面求 I 的最小值．由多元函数求极值的方法可知，$\frac{\partial I}{\partial x_1}=0$，$\frac{\partial I}{\partial x_2}=0$，即

$$2(2x_1 + 4x_2 - 11) \times 2 + 2(3x_1 - 5x_2 - 3) \times 3 + 2(x_1 + 2x_2 - 6) + 2(2x_1 + x_2 - 7) \times 2 = 0,$$

$$2(2x_1 + 4x_2 - 11) \times 4 + 2(3x_1 - 5x_2 - 3) \times (-5) + 2(x_1 + 2x_2 - 6) \times 2 + 2(2x_1 + x_2 - 7) = 0.$$

由此可得

$$x_1 = 3.0403, x_2 = 1.2418.$$

误差的平方和为

$$\delta^2 = (2x_1 + 4x_2 - 11)^2 + (3x_1 - 5x_2 - 3)^2 + (x_1 + 2x_2 - 6)^2 + (2x_1 + x_2 - 7)^2.$$

将 $x_1=3.0403$，$x_2=1.2418$ 带入上式得到 $\delta^2=0.34066$.

事实上，求超定方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 的最小二乘解，也可以转化为直接求正则方程组

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{Ax} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{b}$$

的解即可．在本例中，

$$\boldsymbol{A} = \begin{pmatrix} 2 & 4 \\ 3 & -5 \\ 1 & 2 \\ 2 & 1 \end{pmatrix}, \boldsymbol{b} = \begin{pmatrix} 11 \\ 3 \\ 6 \\ 7 \end{pmatrix}.$$

从而

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} = \begin{pmatrix} 18 & -3 \\ -3 & 46 \end{pmatrix}, \boldsymbol{A}^{\mathrm{T}}\boldsymbol{b} = \begin{pmatrix} 51 \\ 48 \end{pmatrix}.$$

故

$$\begin{pmatrix} 18 & -3 \\ -3 & 46 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 51 \\ 48 \end{pmatrix}.$$

同样可得

$$x_1 = 3.0403, x_2 = 1.2418.$$

上面介绍的最小二乘法有关概念与方法可推广到多元函数．例如已知多元函数

$$y = f(x_1, x_2, \cdots, x_l)$$

的一组测量数据$(x_{1i}, x_{2i}, \cdots, x_{li}, y_i)(i=0,1,2,\cdots,m)$，以及一组权系数 $\omega_i > 0(i=0,1,2,\cdots,m)$．要求函数

$$S(x_1, x_2, \cdots, x_l) = \sum_{k=0}^{n} \alpha_k \varphi_k(x_1, x_2, \cdots, x_l)$$

使得

$$F(a_0, a_1, \cdots, a_n) = \sum_{i=0}^{m} \omega_i [y_i - S(x_{1i}, x_{2i}, \cdots, x_{li})]^2$$

最小．这与前面的极值问题一样，系数 $a_0, a_1, \cdots, a_n$ 同样满足法方程式(4.5)，只是这里

$$(\varphi_k, \varphi_j) = \sum_{i=0}^{m} \omega_i \varphi_k(x_{1i}, x_{2i}, \cdots, x_{li}) \varphi_j(x_{1i}, x_{2i}, \cdots, x_{li}).$$

求解法方程就可得到 $\alpha_k(k=0,1,\cdots,n)$，从而得到 $S(x_1, x_2, \cdots, x_l)$，称为函数 $f(x_1, x_2, \cdots, x_l)$ 的最小二乘拟合．

习　题

1. 已知函数 $y=f(x)$的实测数据表(表 4.5)，试利用最小二乘法求多项式曲线与此数据进行拟合．

表 4.5　实测数据表

x_i	1	2	3	4	6	7	8
y_i	2	3	6	7	5	3	2

2. 用最小二乘法求形如 $y=a+bx^2$ 的多项式，使之与表 4.6 中数据相拟合．

表 4.6　一维数据表

x	19	25	31	38	44
y	19.0	32.3	49.0	73.3	97.8

3. 确定经验公式 $y=ae^{bx}$ 中的参数 a 和 b，使该曲线与表 4.7 中数据相拟合．

表 4.7　一维数据表

x_i	1	2	3	4
y_i	60	30	20	15

4. 求线性方程组的最小二乘解.

$$\begin{bmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 1 \\ 0 \end{bmatrix}$$

5. 给出平面函数 $z(x,y)=ax+by+c$ 的数据见表 4.8.

表 4.8　二维数据表

x_i	0.1	0.2	0.4	0.6	0.9
y_i	0.2	0.3	0.5	0.7	0.8
z_i	0.58	0.63	0.73	0.83	0.92

按最小二乘原理确定 a、b、c.

5　数值积分与数值微分

5.1　引言

5.1.1　数值积分的基本思想

科学和工程计算常常需要计算积分.

例 5.1　数学中一些重要的函数如:

误差函数　　$\operatorname{erf}(x)=\dfrac{2}{\sqrt{\pi}}\int_0^x \mathrm{e}^{-t^2}\,\mathrm{d}t$,

余误差函数　　$\operatorname{erfc}(x)=\dfrac{2}{\sqrt{\pi}}\int_x^{+\infty} \mathrm{e}^{-t^2}\,\mathrm{d}t$,

Γ 函数　　$\Gamma(x)=\int_0^{+\infty} t^{x-1}\mathrm{e}^{-t}\mathrm{d}t\quad(x>0)$,

Bessel 函数　　$\mathrm{J}_v(x)=\dfrac{1}{\sqrt{\pi}\Gamma\left(v+\dfrac{1}{2}\right)}\left(\dfrac{x}{2}\right)^v\int_0^{\pi}\cos(x\cos\theta)\sin^{2v}\theta\mathrm{d}\theta$,

等等都和积分有关系.

例 5.2　在概率论和统计学中的一些重要概念如:

正态分布的分布函数　　$F(x)=\dfrac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x}\mathrm{e}^{\frac{(x-\mu)^2}{2\sigma^2}}\mathrm{d}x$,

连续性随机变量 ξ 的数学期望　　$E\xi=\int_{-\infty}^{+\infty}xp(x)\mathrm{d}(x)$,

等也和积分有关系.

例 5.3　设有均质等厚各向同性的水平储层,液体单相微可压缩,无外来能源供给,储层无限大,只有一口井以恒定产量生产,不考虑井筒储存和表皮因子的影响,这时储层的压力分布满足:

$$\begin{cases}\dfrac{1}{r}\dfrac{\partial}{\partial r}\left(r\dfrac{\partial p}{\partial r}\right)=\dfrac{1}{\eta}\dfrac{\partial p}{\partial t},\\ p(r,0)=p_0,\\ \left(r\dfrac{\partial p}{\partial r}\right)\Big|_{r=r_w}=\dfrac{qB\mu}{2\pi kh},\\ p(\infty,t)=p_0.\end{cases}$$

利用 Laplace 变换可得该定解问题的无因次压力 $P_D(r_D,t_D)$ 的实空间解析解

$$P_D(r_D,t_D)=\frac{2}{\pi}\int_0^{+\infty}\frac{(1-\mathrm{e}^{-u^2t_D})[J_1(u)Y_0(ur_D)-J_0(ur_D)Y_1(u)]}{u^2[J_1^2(u)+Y_1^2(u)]}\mathrm{d}u,$$

这个解也和积分有关系.

在利用有限元方法或者边界元方法求偏微分方程的数值解时,同样需要计算大量的积分.在求积分方程数值解时,也需要计算大量的积分.

还有很多重要的函数、概念和计算方法等都和积分有关系,都需要计算积分.在计算某些积分的时候就会发现,采用 Newton—Leibniz 公式根本行不通,这是因为这些被积函数根本不存在显式的原函数.有时,甚至只知道被积函数在一些离散点上的一些函数值.因此有必要研究积分的数值计算问题.

根据积分中值定理,若函数 $f(x)$在闭区间$[a,b]$上连续,则在闭区间$[a,b]$上必然存在一点 ζ,使

$$\int_a^b f(x)\mathrm{d}x=(b-a)f(\zeta) \tag{5.1}$$

成立.就是说,从几何意义来看,底为 $b-a$ 而高为 $f(\zeta)$的矩形面积恰等于所求曲边梯形的面积.问题在于点 ζ 的具体位置一般是不知道的,因而难以准确算出 $f(\zeta)$的值.将 $f(\zeta)$称为区间$[a,b]$上的平均高度,这样,只要对平均高度 $f(\zeta)$提供一种算法,相应地便获得一种数值积分方法.

例如,如果用区间$[a,b]$两个端点处的值 $f(a)$和 $f(b)$分别作为 $f(\zeta)$的近似值,就可以得到两个求积公式为

$$\int_a^b f(x)\mathrm{d}x\approx(b-a)f(a), \tag{5.2}$$

$$\int_a^b f(x)\mathrm{d}x\approx(b-a)f(b). \tag{5.3}$$

式(5.2)和式(5.3)分别称为左矩形公式和右矩形公式.如果用区间中点的函数值 $f\left(\frac{a+b}{2}\right)$作为 $f(\zeta)$的近似值,则得到中矩形公式

$$\int_a^b f(x)\mathrm{d}x\approx(b-a)f\left(\frac{a+b}{2}\right). \tag{5.4}$$

如果用两端点的"高度"$f(a)$与 $f(b)$取算术平均值作为平均高度 $f(\zeta)$的近似值,则得到如下的梯形公式

$$\int_a^b f(x)\mathrm{d}x\approx\frac{b-a}{2}[f(a)+f(b)]. \tag{5.5}$$

一般地,可以在区间$[a,b]$上适当选取某些节点 x_k,

$$a\leqslant x_0<x_1<\cdots<x_n\leqslant b.$$

然后用 $f(x_k)$加权平均得到平均高度 $f(\zeta)$的近似值,这样构造出的求积公式具有下列形式

$$\int_a^b f(x)\mathrm{d}x\approx\sum_{k=0}^{n}\omega_k f(x_k). \tag{5.6}$$

式中，x_k 称为求积节点，ω_k 称为求积系数，亦称伴随节点 x_k 的权．权 ω_k 仅仅与节点 x_k 的选取有关，而不依赖于被积函数 $f(x)$ 的具体形式．

这类数值积分方法通常称作机械求积，其特点是将积分求值问题归结为函数值的计算，这就避开了牛顿—莱布尼兹公式需要寻求原函数的困难．

建立某类数值求积格式的主要目标就是选择恰当的求积节点 x_k 和权 ω_k，使得求积格式的精度比较高．

5.1.2 代数精度

为了讨论方便，记求积公式(5.6)的左右两端分别为 $I[f]$ 和 $Q[f]$，即

$$I[f]=\int_a^b f(x)\mathrm{d}x, \tag{5.7}$$

$$Q[f]=\sum_{k=0}^{n}\omega_k f(x_k). \tag{5.8}$$

误差余项为

$$R[f]=I[f]-Q[f]. \tag{5.9}$$

数值求积方法是近似方法，要保证精度，就要使求积公式能对“尽可能多”的函数准确地成立，这就提出了所谓代数精度的概念．它是衡量数值积分公式优劣的重要指标之一．

定义 5.1 如果某个求积公式对于次数$\leqslant m$ 的多项式均能准确地成立，但对于 $m+1$ 次多项式就不准确成立，则称该求积公式具有 m 次代数精度．

由于任意次数$\leqslant m$ 的多项式可以由多项式 $1,x,\cdots,x^m$ 唯一线性表示，故某求积公式若具有 m 次代数精度，则必然有

$$I[1]=Q[1],I[x]=Q[x],\cdots,I[x^m]=Q[x^m],I[x^{m+1}]\neq Q[x^{m+1}],$$

即

$$\begin{cases}\omega_0\cdot 1+\omega_1\cdot 1+\cdots+\omega_n\cdot 1=\int_a^b 1\mathrm{d}x=b-a,\\ \omega_0\cdot x_0+\omega_1\cdot x_1+\cdots+\omega_n\cdot x_n=\int_a^b x\mathrm{d}x=(b^2-a^2)/2,\\ \cdots\cdots\\ \omega_0\cdot x_0^m+\omega_1\cdot x_1^m+\cdots+\omega_n\cdot x_n^m=\int_a^b x^m\mathrm{d}x=(b^{m+1}-a^{m+1})/m+1,\end{cases} \tag{5.10}$$

且

$$\omega_0\cdot x_0^{m+1}+\omega_1\cdot x_1^{m+1}+\cdots+\omega_n\cdot x_n^{m+1}\neq\int_a^b x^{m+1}\mathrm{d}x. \tag{5.11}$$

例 5.4 验证左矩形求积公式只有零次代数精度，梯形求积公式具有 1 次代数精度．

解 对于左矩形公式 $Q[f]=f(a)(b-a)$，从而

$$I[1]=b-a,\quad I[x]=(b^2-a^2)/2,$$

$$Q[1]=b-a,\quad Q[x]=a(b-a),$$

从而左矩形公式具有零次代数精度.

对于梯形公式 $Q[f]=\frac{b-a}{2}[f(a)+f(b)]$,从而

$$I[1]=b-a,\quad I[x]=(b^2-a^2)/2,\quad I[x^2]=(b^3-a^3)/3,$$

$$Q[1]=b-a,\quad Q[x]=(b^2-a^2)/2,\quad Q[x^2]=(b-a)(b^2+a^2)/2,$$

从而梯形求积公式具有一次代数精度.

同样可以验证中矩形公式具有 1 次代数精度.

例 5.5 确定下列求积公式中的待定参数,使其精度尽可能高,并指明其代数精度.

(1) $\int_{-h}^{h}f(x)\mathrm{d}x=\omega_{-1}f(-h)+\omega_0 f(0)+\omega_1 f(h)$;

(2) $\int_{-1}^{1}f(x)\mathrm{d}x\approx\frac{1}{3}f(-1)+\frac{2}{3}f(x_1)+f(x_2)$.

解 (1)求积公式含有三个待定参数,故令求积公式对 $f(x)=1,x,x^2$ 准确成立,即

$$\begin{cases}\omega_{-1}+\omega_0+\omega_1=\int_{-h}^{h}1\mathrm{d}x=2h,\\ \omega_{-1}\cdot(-h)+\omega_0\cdot 0+\omega_1\cdot h=\int_{-h}^{h}x\mathrm{d}x=0,\\ \omega_{-1}\cdot(-h)^2+\omega_0\cdot 0+\omega_1\cdot h^2=\int_{-h}^{h}x^2\mathrm{d}x=2h^3/3.\end{cases}$$

从而得到

$$\omega_{-1}=\omega_1=\frac{1}{3}h,\omega_0=\frac{4}{3}h.$$

求出了权重 $\omega_{-1},\omega_0,\omega_1$ 后,再将 $f(x)=x^3,x^4$ 代入计算可得

$$I[x^3]=0,I[x^4]=2h^5/5,$$

$$Q[x^3]=0,Q[x^4]=2h^5/3.$$

从而 $I[x^4]\neq Q[x^4]$,公式具有 3 次代数精度.

(2)求积公式含有两个未知节点 x_1 和 x_2. 当 $f(x)=1$ 时

$$I[1]=2,Q[1]=2.$$

故令求积公式对 $f(x)=x,x^2$ 准确成立,即

$$\begin{cases}2x_1+3x_2=1,\\ 2x_1^2+3x_2^2=1.\end{cases}$$

从而得到

$$\begin{cases} x_1 = 0.68990, \\ x_2 = -0.12660, \end{cases} \quad \text{或} \quad \begin{cases} x_1 = -0.28990, \\ x_2 = 0.52660. \end{cases}$$

求出了节点 x_1 和 x_2 后，再将 $f(x)=x^3$ 代入可得

$$I[x^3] = 0, Q[x^3] = \frac{1}{3}[-1 + 2x_1^3 + 3x_2^3] \neq 0.$$

从而 $I[x^3] \neq Q[x^3]$，求积公式具有 2 次代数精度．

如果事先选定求积节点 x_k，例如 $a \leqslant x_0 < x_1 < \cdots < x_n \leqslant b$[如例 5.5(1)]，这时只要取 $m=n$，这时式(5.10)是关于权 $\omega_0, \omega_1, \cdots, \omega_n$ 的线性代数方程组．由于方程组的系数矩阵的行列式是范德蒙行列式，节点互不相等，从而方程组有唯一解．故求积公式(5.6)至少具有 n 次代数精度．但是如果求积节点太多，方程组的求解还是比较麻烦．下节中将通过另一种简单办法即利用 Lagrange 插值来确定权 ω_k.

如果求积节点 x_k 和权 ω_k 都未知[特别是 x_k 未知如例 5.5(2)]，则式(5.10)一般是非线性代数方程组，求解非常困难．后面将借助于正交多项式先求出 x_k，再确定权 ω_k，这就是后面将要介绍的 Gauss 型求积公式．

5.1.3 插值型的求积公式

设给定一组节点

$$a \leqslant x_0 < x_1 < x_2 < \cdots < x_n \leqslant b$$

且已知函数 $f(x)$ 在这些节点上的值，作插值函数

$$L_n(x) = \sum_{k=0}^{n} f(x_k) l_k(x).$$

由于代数多项式 $L_n(x)$ 的原函数是容易求出的，取

$$Q[f] = \int_a^b L_n(x)\mathrm{d}x$$

作为积分 $I[f] = \int_a^b f(x)\mathrm{d}x$ 的近似值，这样构造出的求积公式

$$Q[f] = \int_a^b \sum_{k=0}^{n} f(x_k) l_k(x)\mathrm{d}x = \sum_{k=0}^{n} \left[f(x_k) \int_a^b l_k(x)\mathrm{d}x \right] = \sum_{k=0}^{n} \omega_k f(x_k) \tag{5.12}$$

称作是插值型的，式中求积系数 ω_k 通过插值基函数 $l_k(x)$ 积分得出

$$\omega_k = \int_a^b l_k(x)\mathrm{d}x. \tag{5.13}$$

由插值余项定理即知，对于插值型的求积公式(5.12)，其误差余项为

$$R[f] = I[f] - Q[f] = \int_a^b \frac{f^{(n+1)}(\xi)}{(n+1)!} \omega_{n+1}(x)\mathrm{d}x. \tag{5.14}$$

式中 ξ 与变量 x 有关，$\omega_{n+1}(x) = (x-x_0)(x-x_1)\cdots(x-x_n)$.

如果求积公式(5.12)是插值型的，按式(5.14)，对于次数$\leqslant n$的多项式$f(x)$，其余项$R[f]$等于0，因而这时求积公式至少具有n次代数精度.

反之，如果求积公式(5.12)至少具有n次代数精度，则它必定是插值型的. 事实上，这时式(5.12)对于插值基函数$l_k(x)$应准确成立，即有

$$\int_a^b l_k(x)\mathrm{d}x = \sum_{j=0}^{n} \omega_j l_k(x_j).$$

注意到以$l_k(x_j)=\delta_{kj}$，上式右端实际上等于ω_k，因而式(5.13)成立.

综上所述，可以得到：

定理 5.1 形如式(5.12)的求积公式至少有n次代数精度的充分必要条件是，它是插值型的.

特别地，当$f(x)=1$时，带入插值型求积公式(5.12)，由余项为零可得

$$\sum_{k=0}^{n} \omega_k = b-a. \tag{5.15}$$

5.1.4 求积公式的稳定性

在求积公式(5.6)中，由于计算$f(x_k)$可能产生误差δ_k，实际上得到了$\tilde{f}_k$，即$f(x_k)=\tilde{f}_k+\delta_k$. 记

$$Q[f] = \sum_{k=0}^{n} \omega_k f(x_k), Q[\tilde{f}] = \sum_{k=0}^{n} \omega_k \tilde{f}_k.$$

如果对$\forall \varepsilon>0$，只要误差$|\delta_k|$充分小，就有

$$\left| Q[f]-Q[\tilde{f}] \right| \leqslant \varepsilon, \tag{5.16}$$

就表明求积公式(5.6)计算是稳定的. 于是给出下面的定义.

定义 5.2 对于$\forall \varepsilon>0$，存在$\delta>0$，当$\left| f(x_k)-\tilde{f}_k \right| \leqslant \delta (k=0,1,2,\cdots,n)$，就有式(5.16)成立，则称求积公式(5.6)是稳定的.

定理 5.2 若求积公式(5.6)中系数$\omega_k>0 (k=0,1,2,\cdots,n)$，则该求积公式稳定.

证明从略. 定理5.2表明只要所有的求积系数$\omega_k>0$，就能保证计算的稳定性.

5.2 牛顿—柯特斯公式

5.2.1 牛顿—柯特斯公式的内容

将积分区间$[a,b]n$等分，步长$h=\dfrac{b-a}{n}$，选取等距节点

$$x_k = a+kh \quad (k=0,1,2,\cdots,n).$$

构造出的插值型求积公式

$$Q[f]=(b-a)\sum_{k=0}^{n}C_k^{(n)}f(x_k) \tag{5.17}$$

称作牛顿—柯特斯(Newton—Cotes)公式,式中 $C_k^{(n)}$ 称作柯特斯系数. 引进变换 $x=a+th$,由式(5.12)和式(5.13)得

$$C_k^{(n)}=\frac{\omega_k}{b-a}=\frac{1}{b-a}\int_a^b l_k(x)\mathrm{d}x=\frac{1}{b-a}\int_a^b\prod_{\substack{j=0\\j\neq k}}^{n}\frac{x-x_j}{x_k-x_j}\mathrm{d}x$$

$$=\frac{h}{b-a}\int_0^n\prod_{\substack{j=0\\j\neq k}}^{n}\frac{t-j}{k-j}\mathrm{d}t=\frac{(-1)^{n-k}}{nk!(n-k)!}\int_0^n\prod_{\substack{j=0\\j\neq k}}^{n}(t-j)\mathrm{d}t. \tag{5.18}$$

从式(5.18)可以看出,柯特斯系数 $C_k^{(n)}$ 与积分区间和被积函数没有关系. 因此,只要事先计算出 $C_k^{(n)}$ 的值,则对任意的积分区间$[a,b]$及被积函数 $f(x)$,代入式(5.17)即可计算出积分的近似值.

由于式(5.18)是多项式的积分,柯特斯系数的计算不会遇到实质性的困难. 关于柯特斯系数 $C_k^{(n)}$,有如下两个性质:

(1)$\sum_{k=0}^{n}C_k^{(n)}=1$;

(2)$C_k^{(n)}=C_{n-k}^{(n)}$.

下面给出几个常见的求积公式:

(1)当 $n=1$ 时,$C_0^{(1)}=C_1^{(1)}=\frac{1}{2}$,相应的求积公式为

$$\int_a^b f(x)\mathrm{d}x\approx\frac{b-a}{2}[f(a)+f(b)]=T, \tag{5.19}$$

该求积公式是梯形公式(5.5).

(2)当 $n=2$ 时,由式(5.18)得柯特斯系数为

$$C_0^{(2)}=\frac{1}{4}\int_0^2(t-1)(t-2)\mathrm{d}t=\frac{1}{6}$$

$$C_1^{(2)}=-\frac{1}{2}\int_0^2 t(t-2)\mathrm{d}t=\frac{4}{6}$$

$$C_2^{(2)}=\frac{1}{4}\int_0^2(t-1)\mathrm{d}t=\frac{1}{6}$$

相应的求积公式是下列辛普森(Simpson)公式(或称为抛物线求积公式)

$$\int_a^b f(x)\mathrm{d}x\approx\frac{b-a}{6}\left[f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right]=S. \tag{5.20}$$

(3)当 $n=4$ 时,有

$$\int_a^b f(x)\mathrm{d}x\approx\frac{b-a}{90}[7f(x_0)+32f(x_1)+12f(x_2)+32f(x_3)+7f(x_4)]=C, \tag{5.21}$$

称为柯特斯公式,其中

$$x_k = a + kh, h = \frac{b-a}{4} \quad (k = 0,1,2,3,4).$$

表 5.1 列出柯特斯系数表开头的一部分.

表 5.1 柯特斯系数

n	$C_k^{(n)}$								
1	$\frac{1}{2}$	$\frac{1}{2}$							
2	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$						
3	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	$\frac{3}{8}$					
4	$\frac{7}{90}$	$\frac{16}{45}$	$\frac{2}{15}$	$\frac{16}{45}$	$\frac{7}{90}$				
5	$\frac{19}{288}$	$\frac{25}{96}$	$\frac{25}{144}$	$\frac{25}{144}$	$\frac{25}{96}$	$\frac{19}{288}$			
6	$\frac{41}{840}$	$\frac{9}{35}$	$\frac{9}{280}$	$\frac{34}{105}$	$\frac{9}{280}$	$\frac{9}{35}$	$\frac{41}{840}$		
7	$\frac{751}{17280}$	$\frac{3577}{17280}$	$\frac{1323}{17280}$	$\frac{2989}{17280}$	$\frac{2989}{17280}$	$\frac{1323}{17280}$	$\frac{3577}{17280}$	$\frac{751}{17280}$	
8	$\frac{989}{28350}$	$\frac{5888}{28350}$	$\frac{-928}{28350}$	$\frac{10496}{28350}$	$\frac{-4540}{28350}$	$\frac{10496}{28350}$	$\frac{-928}{28350}$	$\frac{5888}{28350}$	$\frac{989}{28350}$

(4)当 $n=8$ 时,柯特斯系数有正有负(实际上 $n\geqslant 8$ 时也是如此.),这时稳定性得不到保证. 因此,实际计算不用高阶的牛顿—柯特斯公式.

牛顿—柯特斯公式作为插值型求积公式,它至少有 n 次代数精度. 但是由于这种插值是等距节点插值,它的代数精度还可以进一步提高. 下面先看一个例子.

例 5.6 验证 $n=2$ 的辛普森公式具有 3 次代数精度.

解 根据 $I[f] = \int_a^b f(x)\mathrm{d}x, Q[f] = S = \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]$, 经过计算可得

$$I[1] = Q[1] = b-a,$$

$$I[x] = Q[x] = (b^2 - a^2)/2,$$

$$I[x^2] = Q[x^2] = (b^3 - a^3)/3,$$

$$I[x^3] = Q[x^3] = (b^4 - a^4)/4,$$

$$I[x^4] = (b^5 - a^5)/5,$$

$$Q[x^4] = (b^5 - a^5)/6 + (b-a)(a^4 + 2a^2b^2 + b^4)/24.$$

显然 $I[x^4] \neq Q[x^4]$,由代数精度的定义可知辛普森求积公式具有 3 次代数精度.

一般地有下述结果：

定理 5.3 当 n 为偶数时，牛顿—柯特斯公式(5.17)至少有 $n+1$ 次代数精度.

5.2.2 几个常用的牛顿—柯特斯公式的余项

5.2.2.1 梯形公式的余项

按余项公式(5.14)，梯形公式(5.19)的余项为

$$R = I[f]-Q[f] = I[f]-T = \int_a^b \frac{f''(\xi)}{2}(x-a)(x-b)\mathrm{d}x,$$

其中 $\xi\in(a,b)$ 且依赖于 x. 设 $f''(x)$ 在 $[a,b]$ 上连续，由于 $(x-a)(x-b)$ 在区间 $[a,b]$ 上不变号(非正)，应用积分中值定理可知，存在点 $\eta\in[a,b]$，使

$$\begin{aligned} R_T &= \frac{f''(\eta)}{2}\int_a^b (x-a)(x-b)\mathrm{d}x \\ &= -\frac{f''(\eta)}{12}(b-a)^3, \eta\in[a,b]. \end{aligned} \tag{5.22}$$

5.2.2.2 辛普森公式的余项

由余项公式(5.14)可得辛普森公式的余项为

$$R = I[f]-Q[f] = I[f]-S = \int_a^b \frac{f'''(\xi)}{3!}(x-a)\left(x-\frac{a+b}{2}\right)(x-b)\mathrm{d}x.$$

由于此时 $(x-a)\left(x-\frac{a+b}{2}\right)(x-b)$ 在 $[a,b]$ 上变号，所以不能直接应用积分中值定理来讨论. 为此构造次数$\leqslant$3 的多项式 $H(x)$，使满足

$$\begin{aligned} &H(a) = f(a), H(b) = f(b), \\ &H\left(\frac{a+b}{2}\right) = f\left(\frac{a+b}{2}\right), H'\left(\frac{a+b}{2}\right) = f'\left(\frac{a+b}{2}\right). \end{aligned} \tag{5.23}$$

由于辛普森公式具有 3 次代数精度，它对于这样构造出的三次多项式 $H(x)$ 是准确的：

$$\begin{aligned} \int_a^b H(x)\mathrm{d}x &= \frac{b-a}{6}\left[H(a)+4H\left(\frac{a+b}{2}\right)+H(b)\right] \\ &= \frac{b-a}{6}\left[f(a)+4f\left(\frac{a+b}{2}\right)+f(b)\right] = S. \end{aligned}$$

因此积分余项为

$$R_S = I-S = \int_a^b [f(x)-H(x)]\mathrm{d}x.$$

对于满足条件式(5.23)的多项式 $H(x)$，由例 3.3 知其插值余项为

$$f(x)-H(x) = \frac{f^{(4)}(\xi)}{4!}(x-a)\left(x-\frac{a+b}{2}\right)^2(x-b),$$

故有

$$R_S = \int_a^b \frac{f^{(4)}(\xi)}{4!}(x-a)\left(x-\frac{a+b}{2}\right)^2(x-b)\mathrm{d}x.$$

这时函数$(x-a)\left(x-\frac{a+b}{2}\right)^2(x-b)$在$[a,b]$上不变号(非正),用积分中值定理有

$$\begin{aligned} R_S &= \frac{f^{(4)}(\eta)}{4!}\int_a^b (x-a)\left(x-\frac{a+b}{2}\right)^2(x-b)\mathrm{d}x \\ &= -\frac{b-a}{180}\left(\frac{b-a}{2}\right)^4 f^{(4)}(\eta). \end{aligned} \tag{5.24}$$

5.2.2.3 柯特斯公式的余项

关于柯特斯公式(5.21)的积分余项,这里不再具体推导,仅列出结果如下:

$$R_C = I - C = -\frac{2(b-a)}{945}\left(\frac{b-a}{4}\right)^6 f^{(6)}(\eta). \tag{5.25}$$

例 5.7 分别利用中矩形公式、梯形公式和辛普森公式计算积分 $I[f]=\int_0^1 \mathrm{e}^{-x^2}\mathrm{d}x$.

解

$$M = (1-0)\exp(-0.5) \approx 0.778801,$$

$$T = 0.5(\exp(0)+\exp(-1)) \approx 0.683940,$$

$$S = \frac{1}{6}(\exp(0)+4\exp(-0.25)+\exp(-1)) \approx 0.747180.$$

例 5.8 分别利用梯形公式和 Simpson 公式计算 $I[f]=\int_0^1 \mathrm{e}^{-x}\mathrm{d}x$,并估计误差界.

解

$$T = 0.5\times[\exp(0)+\exp(-1)] \approx 0.683940,$$

$$R_T(f) = -\frac{(b-a)^3}{12}f''(\eta) = -\frac{1}{12}\mathrm{e}^{-\eta} \quad \eta\in[0,1],$$

$$|R_T(f)| \leqslant \frac{1}{12} = 0.083333,$$

$$S = \frac{1}{6}\times[\exp(0)+4\exp(-0.5)+\exp(-1)] \approx 0.6323337,$$

$$R_S(f) = -\frac{(b-a)^5}{2880}f^{(4)}(\eta) = -\frac{1}{2880}\mathrm{e}^{-\eta} \quad \eta\in[0,1],$$

$$|R_S(f)| \leqslant \frac{1}{2880} = 0.0003472.$$

例 5.9 利用数值积分法求解积分方程

$$\varphi(x)+\int_0^1 x(\mathrm{e}^{xt}-1)\varphi(t)\mathrm{d}t = \mathrm{e}^x - x \tag{5.26}$$

解 通常把积分号下含有未知函数的方程称为积分方程. 根据积分方程的分类,式

(5.26)称为一维线性的第二类 Fredholm 方程．一些简单的积分方程可以求出其准确解，大多数方程只能求出其近似解．下面说明利用数值积分的方法近似求解积分方程的过程．考虑第二类 Fredholm 积分方程

$$\varphi(x)-\lambda\int_a^b K(x,t)\varphi(t)\mathrm{d}t=f(x), \tag{5.27}$$

任取一种数值积分公式(5.6)近似代替式(5.27)中的积分，得

$$\varphi(x)-\lambda\sum_{k=0}^{n}\omega_k K(x,x_k)\varphi(x_k)\approx f(x).$$

再令 $x=x_j(j=0,1,2,\cdots,n)$，得

$$\varphi(x_j)-\lambda\sum_{k=0}^{n}\omega_k K(x_j,x_k)\varphi(x_k)\approx f(x_j).$$

考虑代数方程组

$$\tilde{\varphi}(x_j)-\lambda\sum_{k=0}^{n}\omega_k K(x_j,x_k)\tilde{\varphi}(x_k)=f(x_j), \tag{5.28}$$

这是含有 $n+1$ 个未知数 $\tilde{\varphi}(x_0),\tilde{\varphi}(x_1),\cdots,\tilde{\varphi}(x_n)$ 的线性代数方程组．若求得其解，则可作为 $\varphi(x)$ 在节点 $x_0,x_1,\cdots,x_n$ 的近似值，从而可取积分方程式(5.27)的近似解为

$$\tilde{\varphi}(x)=\lambda\sum_{k=0}^{n}\omega_k K(x,x_k)\tilde{\varphi}(x_k)+f(x). \tag{5.29}$$

根据上述思路，在式(5.26)中，取节点为 $x_0=0,x_1=0.5,x_2=1$．令式(5.26)中的 x 分别取值 0，0.5 和 1，并采用辛普森求积公式来计算式(5.26)中的积分，得

$$\varphi(0)=1,$$

$$\frac{\mathrm{e}^{0.25}+2}{3}\varphi(0.5)+\frac{\mathrm{e}^{0.5}-1}{12}\varphi(1)=\mathrm{e}^{0.5}-0.5,$$

$$\frac{2(\mathrm{e}^{0.5}-1)}{3}\varphi(0.5)+\frac{\mathrm{e}+5}{6}\varphi(1)=\mathrm{e}-1.$$

从而得到

$$\begin{cases}\varphi(0)=1,\\ 1.0947\varphi(0.5)+0.0541\varphi(1)=1.1487,\\ 0.4325\varphi(0.5)+1.2864\varphi(1)=1.7183.\end{cases}$$

解线性方程组得到

$$\varphi(0)=1,\varphi(0.5)=0.9999,\varphi(1)=0.9996.$$

由式(5.29)可得方程式(5.26)的近似解为

$$\tilde{\varphi}(x)=\mathrm{e}^x-x(0.6666\mathrm{e}^{0.5x}+0.1666\mathrm{e}^x)-0.1668x.$$

事实上，式(5.26)的精确解为 $\varphi(x)=1$.

5.3 复化求积公式

前已指出，使用高阶牛顿—柯特斯公式不能保证计算的稳定性．故一般不采用提高阶(缩小步长)的办法．为了改善求积的精度，通常采用两种途径：一是复化求积法，二是采用非等距节点(后面将要讲到的高斯方法)．下面先介绍复化求积公式．所谓复化求积法，就是先用低阶的牛顿—柯特斯公式求得每个子区间$[x_k,x_{k+1}]$上的积分值I_k，然后再求和，用$\sum_{k=0}^{n-1}I_k$作为所求积分I的近似值．

5.3.1 复化梯形公式

将积分区间$[a,b]$划为n等分，步长$h=\dfrac{b-a}{n}$，分点$x_k=a+kh(k=0,1,\cdots,n)$．在每个子区间上$[x_k,x_{k+1}](k=0,1,\cdots,n-1)$采用梯形公式(5.19)，得到

$$I[f]=\int_a^b f(x)\mathrm{d}x=\sum_{k=0}^{n-1}\int_{x_k}^{x_{k+1}}f(x)\mathrm{d}x=\sum_{k=0}^{n-1}\frac{h}{2}[f(x_k)+f(x_{k+1})]+R_n(f).$$

记

$$T_n=\sum_{k=0}^{n-1}\frac{h}{2}[f(x_k)+f(x_{k+1})]=\frac{h}{2}[f(a)+2\sum_{k=1}^{n-1}f(x_k)+f(b)],\tag{5.30}$$

称为复化梯形公式，其余项可由式(5.22)得

$$R_n(f)=I-T_n=\sum_{k=0}^{n-1}\left[-\frac{h^3}{12}f''(\eta_k)\right]\quad \eta_k\in[x_k,x_{k+1}].$$

设$f(x)\in C^2[a,b]$，且

$$\min_{0\leqslant k\leqslant n-1}f''(\eta_k)\leqslant\frac{1}{n}\sum_{k=0}^{n-1}f''(\eta_k)\leqslant\max_{0\leqslant k\leqslant n-1}f''(\eta_k),$$

故存在$\eta\in[a,b]$使得

$$f''(\eta)=\frac{1}{n}\sum_{k=0}^{n-1}f''(\eta_k).$$

于是复化梯形公式的余项为

$$R_n(f)=\sum_{k=0}^{n-1}\left[-\frac{h^3}{12}f''(\eta_k)\right]=-\frac{nh^3}{12}f''(\eta)=-\frac{b-a}{12}h^2f''(\eta).\tag{5.31}$$

5.3.2 复化辛普森公式

将积分区间$[a,b]$划为n等分，步长$h=\dfrac{b-a}{n}$，分点$x_k=a+kh(k=0,1,\cdots,n)$．在每个子区间上$[x_k,x_{k+1}](k=0,1,\cdots,n-1)$采用辛普森公式(5.20)，若记$[x_k,x_{k+1}]$的中点为$x_{k+1/2}=$

$x_k+\frac{1}{2}h$，得到

$$I[f]=\int_a^b f(x)\mathrm{d}x=\sum_{k=0}^{n-1}\int_{x_k}^{x_{k+1}} f(x)\mathrm{d}x$$

$$=\sum_{k=0}^{n-1}\frac{h}{6}[f(x_k)+4f(x_{k+1/2})+f(x_{k+1})]+R_n(f).$$

记

$$S_n=\sum_{k=0}^{n-1}\frac{h}{6}[f(x_k)+4f(x_{k+1/2})+f(x_{k+1})]$$

$$=\frac{h}{6}\Big[f(a)+4\sum_{k=0}^{n-1}f(x_{k+1/2})+2\sum_{k=1}^{n-1}f(x_k)+f(b)\Big],\tag{5.32}$$

称为复化辛普森公式．其余项由式(5.24)得

$$R_n(f)=I-S_n=\sum_{k=0}^{n-1}\Big[-\frac{h^5}{2880}f^{(4)}(\eta_k)\Big]\quad \eta_k\in[x_k,x_{k+1}].$$

设 $f(x)\in C^4[a,b]$，与复化梯形公式类似可得

$$R_n(f)=-\frac{b-a}{2880}h^4 f^{(4)}(\eta)\quad \eta\in[a,b].\tag{5.33}$$

注意到 T_n、S_n 中的求积系数都是大于零的，由定理 5.2 可知，复化梯形公式和复化辛普森公式是稳定的．

其他牛顿—柯特斯公式亦可用类似的手续加以复化．

例 5.10 用两种方法计算积分 $\int_0^1\frac{\sin x}{x}\mathrm{d}x$.

(1)复化梯形公式($n=8$)；

(2)复化辛普森公式($n=4$).

解 将区间分为 8 等分，先计算出各个节点处的函数值．令 $f(x)=\frac{\sin x}{x}$，则有

$$f(0)=1,f(1/8)=0.9973976,f(2/8)=0.9896158,$$

$$f(3/8)=0.9767267,f(4/8)=0.9585510,f(5/8)=0.9361556,$$

$$f(6/8)=0.9088516,f(7/8)=0.8771925,f(1)=0.8414709.$$

(1)区间 8 等分时，$h=1/8$，由式(5.30)可知

$$T_8=\frac{1}{16}\times\{f(0)+2[f(1/8)+f(2/8)+f(3/8)+f(4/8)$$

$$+f(5/8)+f(6/8)+f(7/8)+f(1)]\}$$

$$=0.9456909.$$

(2)区间 4 等分时，$h=1/4$，由式(5.32)可知

$$\begin{aligned} S_4 &= \frac{1}{24}\times\{f(0)+4[f(1/8)+f(3/8)+f(5/8)+f(7/8)] \\ &\quad +2\times[f(2/8)+f(4/8)+f(6/8)+f(1)]\} \\ &= 0.9460832. \end{aligned}$$

比较上面两个结果 T_8 与 S_4，它们都需要提供 9 个点上的函数值，计算量基本相同，然而精度却差别很大，同积分的准确值 $I=0.9460831$ 比较，复化梯形法的结果 $T_8=0.9456909$ 只有 2 位有效数字，而复化辛普森公式却有 6 位有效数字．

5.4 龙贝格求积公式

5.4.1 梯形法的递推化

上一节介绍的复化求积方法对提高精度是行之有效的，但在使用求积公式之前必须给出合适的步长，步长取得太大精度难以保证，步长太小则会导致计算量的增加，而事先给出一个恰当的步长又往往是困难的．实际计算时若精度不够可将步长逐次分半．设将区间$[a,b]$分为 n 等分，共有 $n+1$ 个分点，如果将求积区间再二分 1 次，则分点增加为 $2n+1$ 个．将二分前后两个积分值联系起来加以考察．注意到每个子区间$[x_k,x_{k+1}]$经过二分只增加了一个分点 $x_{k+1/2}=\frac{1}{2}(x_k+x_{k+1})$，用复化梯形公式求得该子区间上的积分值为

$$\frac{h}{4}[f(x_k)+2f(x_{k+1/2})+f(x_{k+1})].$$

注意，这里 $h=\frac{b-a}{n}$代表二分前的步长．将每个子区间上的积分值相加得

$$T_{2n}=\frac{h}{4}\sum_{k=0}^{n-1}[f(x_k)+f(x_{k+1})]+\frac{h}{2}\sum_{k=0}^{n-1}f(x_{k+\frac{1}{2}}),$$

从而利用式(5.30)可导出下列递推公式

$$T_{2n}=\frac{1}{2}T_n+\frac{h}{2}\sum_{k=0}^{n-1}f(x_{k+\frac{1}{2}}). \tag{5.34}$$

从而可知，当计算 T_{2n}时，如果利用前一次的计算结果 T_n，则仅仅需要计算二分后新增加的 n 个节点 $x_{k+1/2}$上的函数值．由此可见式(5.34)避免了“老节点”上的函数值的重复计算，从而节约了大致一半的计算量．

例 5.11 计算积分值 $I=\int_0^1\frac{\sin x}{x}\mathrm{d}x$ 的近似值，要求计算精度满足

$$|T_{2n}-T_n|\leqslant 10^{-7}.$$

解 先对整个区间$[0,1]$使用梯形公式．对于函数 $f(x)=\frac{\sin x}{x}$，它在 $x=0$ 的值定义为

$f(0)=1$，而 $f(1)=0.8414709$，据梯形公式计算得

$$T_1=\frac{1}{2}[f(0)+f(1)]=0.9207355.$$

然后将区间二等分，再求出中点的函数值 $f\left(\frac{1}{2}\right)=0.9588510$，从而利用递推公式(5.34)，有

$$T_2=\frac{1}{2}T_1+\frac{1}{2}f\left(\frac{1}{2}\right)=0.9397933.$$

进一步二分求积区间，并计算新分点上的函数值

$$f(1/4)=0.9896158, f(3/4)=0.9088516.$$

再利用式(5.34)，有

$$T_4=\frac{1}{2}T_2+\frac{1}{4}\left[f\left(\frac{1}{4}\right)+f\left(\frac{3}{4}\right)\right]=0.9445135.$$

这样不断二分下去，计算结果见表 5.2(表中 k 代表二分次数，区间等分数 $n=2^k$).

表 5.2　次数与梯形法迭代值

k	T_n	k	T_n
1	0.9397933	6	0.9460769
2	0.9445135	7	0.9460815
3	9.9456909	8	0.9460827
4	0.9459850	9	0.9460830
5	0.9460596	10	0.9460831

从表 5.2 中可以看出，将积分区间二等分了 10 次，求得积分的近似值为 0.9460831 满足精度要求

$$|T_{2^{10}}-T_{2^9}|\leqslant 10^{-7}.$$

这时二等分区间 10 次，用到的分点高达 1025 个，计算量很大.

5.4.2　龙贝格求积公式的内容

梯形法的算法简单，但精度较差，收敛的速度缓慢. 如何提高收敛速度以节省计算量，自然是人们极为关心的问题.

根据梯形法的误差公式(5.22)

$$I-T_n=-\frac{b-a}{12}h^2f''(\eta)\quad \eta\in[a,b],$$

$$I-T_{2n}=-\frac{b-a}{12}\left(\frac{h}{2}\right)^2f''(\bar{\eta})\quad \bar{\eta}\in[a,b],$$

假定 $f''(\eta)\approx f''(\bar{\eta})$，则有

$$\frac{I-T_{2n}}{I-T_n}\approx\frac{1}{4}.$$

将上式移项整理，可得

$$1-T_{2n}\approx\frac{1}{3}(T_{2n}-T_n). \tag{5.35}$$

由此可见，只要二分前后的两个积分值 T_n 与 T_{2n} 相当接近，就可以保证计算结果 T_{2n} 的误差很小．这样直接用计算结果来估计误差的方法通常称作误差的事后估计法．

按式(5.35)，积分近似值 T_{2n} 的误差大致等于 $\frac{1}{3}(T_{2n}-T_n)$．因此如果用这个误差值作为 T_{2n} 的一种补偿，可以期望所得到的

$$\overline{T}=T_{2n}+\frac{1}{3}(T_{2n}-T_n)=\frac{4}{3}T_{2n}-\frac{1}{3}T_n \tag{5.36}$$

可能是更好的结果．

再考察例 5.11，所求得的两个梯形值 $T_4=0.9445135$ 和 $T_8=0.9456909$ 的精度都很差(与准确值 $I=0.9460831$ 比较，只有二三位有效数字)，但如果将它们按式(5.36)作线性组合，则新的近似值

$$\overline{T}=\frac{4}{3}T_{2n}-\frac{1}{3}T_n=0.9460833$$

却有 6 位有效数字．

按式(5.36)组合得到的近似值 $\overline{T}$，其实质究竟是什么呢？直接验证易知

$$S_n=\frac{4}{3}T_{2n}-\frac{1}{3}T_n. \tag{5.37}$$

这就是说，用梯形法二分前后的两个积分值 T_n 与 T_{2n}，按式(5.37)作线性组合，结果得到辛普森法的积分值 S_n．

再考察辛普森法，按误差公式(5.24)，其截断误差大致与 h^4 成正比，因此，若将步长折半则误差将减至原有误差的 1/16，即有

$$\frac{I-S_{2n}}{I-S_n}\approx\frac{1}{16}.$$

由此可得

$$I\approx\frac{16}{15}S_{2n}-\frac{1}{15}S_n.$$

不难直接验证，上式右端的值其实等于 C_n，就是说，用辛普森法二分前后的两个积分值 S_n 与 S_{2n} 按上式作线性组合，结果得到柯特斯法的积分值 C_n，即有

$$C_n\approx\frac{16}{15}S_{2n}-\frac{1}{15}S_n. \tag{5.38}$$

重复同样的手续，依据柯特斯法的误差阶为 h^6，可进一步导出下列龙贝格(Romberg)公式，即

$$R_n=\frac{64}{63}C_{2n}-\frac{1}{63}C_n. \tag{5.39}$$

在变步长的过程中运用式(5.37)、(5.38)和(5.39)，就能将粗糙的梯形值 T_n 逐步加工成

精度较高的辛普森值、柯特斯值和龙贝格值.

计算中经常将龙贝格公式的计算结果列表(表 5.3).

表 5.3　龙贝格迭代过程

k	T_{2^k}	$S_{2^{k-1}}$	$C_{2^{k-2}}$	$R_{2^{k-3}}$
0	T_1			
1	T_2	S_1		
2	T_4	S_2	C_1	
3	T_8	S_4	C_2	R_1
…	…	…	…	…

例 5.12　用龙贝格公式计算 $I=\int_0^1 \frac{\sin x}{x}\mathrm{d}x$.

解　列表计算见表 5.4.

表 5.4　龙贝格迭代值

k	T_{2^k}	$S_{2^{k-1}}$	$C_{2^{k-2}}$	$R_{2^{k-3}}$
0	0.9207355			
1	0.9397933	0.9461459		
2	0.9445135	0.9460869	0.9460830	
3	0.9456909	0.9460833	0.9460831	0.9460831

这里利用二分 3 次的数据(它们的精度都很差,只有二三位是有效数字),通过三次加速求得 $R_1=0.9460831$,这个结果的每一位数字都是有效数字,可见加速的效果是十分显著的.

5.5　高斯公式

5.5.1　Gauss 求积公式的一般理论

形如式(5.6)的机械求积公式

$$\int_a^b f(x)\mathrm{d}x \approx \sum_{k=0}^{n}\omega_k f(x_k) \tag{5.40}$$

含有 $2n+2$ 个待定参数 x_k 和 $\omega_k(k=0,1,\cdots,n)$. 当 x_k 为等距节点时得到的插值型求积公式的代数精度至少为 n 次,如果适当选择这些参数 x_k 和 ω_k,有可能使求积公式具有 $2n+1$ 代数精度. 这类求积公式称为高斯公式. 为使问题更具有一般性,研究带权积分

$$I=\int_a^b \rho(x)f(x)\mathrm{d}x.$$

这里 $\rho(x)$为权函数(见附录 B),类似式(5.40),它的求积公式为

$$\int_a^b \rho(x)f(x)\mathrm{d}x \approx \sum_{k=0}^{n}\omega_k f(x_k). \tag{5.41}$$

注意:前面讲到的插值型求积公式中节点 x_k 时预先给定的,而在式(5.41)中参数 x_k 和 ω_k 都是未知的. 希望能在节点个数固定的情况下,找到恰当的参数,使公式的求积精度最高

引理 5.1 具有 $n+1$ 个节点的求积公式(5.41)的代数精度不超过 $2n+1$ 次.

证明 采用反证法. 只要取一个 $2n+2$ 次多项式让式(5.41)不能准确成立成立即可.

令

$$f(x) = (x-x_0)^2(x-x_1)^2\cdots(x-x_n)^2,$$

显然有

$$\sum_{k=0}^{n}\omega_k f(x_k) = 0, \int_a^b \rho(x)f(x)\mathrm{d}x > 0,$$

从而命题得证.

定义 5.3 如果求积公式(5.41)具有 $2n+1$ 次代数精度,则称其节点 $x_k(k=0,1,\cdots,n)$ 是高斯点,相应的求积公式称为高斯求积公式.

根据定义,要使式(5.41)具有 $2n+1$ 次代数精度,只要取 $f(x)=x^m$,对 $m=0,1,\cdots,2n+1$,式(5.41)都准确成立,即

$$\sum_{k=0}^{n}\omega_k f(x_k) = \int_a^b \rho(x)f(x)\mathrm{d}x \quad (m=0,1,\cdots,2n+1). \tag{5.42}$$

当给定权函数 $\rho(x)$,求出右端积分,则可由式(5.42)解出 x_k 和 ω_k.

例 5.13 试构造下列积分的高斯求积公式

$$\int_0^1 \sqrt{x}f(x)\mathrm{d}x \approx \omega_0 f(x_0) + \omega_1 f(x_1). \tag{5.43}$$

解 令式(5.42)对于 $f(x)=1,x,x^2,x^3$ 准确成立,得到

$$\begin{cases} \omega_0 + \omega_1 = \dfrac{2}{3}, \\ x_0\omega_0 + x_1\omega_1 = \dfrac{2}{5}, \\ x_0^2\omega_0 + x_1^2\omega_1 = \dfrac{2}{7}, \\ x_0^3\omega_0 + x_1^3\omega_1 = \dfrac{2}{9}, \end{cases}$$

这是一个关于未知数 $x_0,x_1,\omega_0,\omega_1$ 的非线性方程组. 经过复杂的演算可得

$$x_0 = 0.821162, x_1 = 0.289949,$$

$$\omega_0 = 0.3889111, \omega_1 = 0.277556.$$

从而形如式(5.43)的高斯公式为

$$\int_0^1 \sqrt{x}f(x)\mathrm{d}x \approx 0.389111f(0.821162) + 0.277556f(0.289949).$$

从此例可以看出求解非线性方程组式(5.43)一般是非常复杂的,通常 $n\geqslant 2$ 就很难求解.

故一般不通过求解方程组式(5.43)来求高斯点 x_k 和权 ω_k，而从分析高斯点的特性入手构造高斯求积公式.

定理 5.2 对于插值型求积公式(5.41)的节点 $a\leqslant x_0<x_1<\cdots<x_n\leqslant b$ 是高斯点的充分必要条件是以这些节点为零点的多项式

$$\omega_{n+1}(x)=(x-x_0)(x-x_1)\cdots(x-x_n)$$

与任何次数不超过 n 的多项式 $P(x)$ 带权 $\rho(x)$ 正交(正交的概念见附录)，即

$$\int_a^b\rho(x)P(x)\omega_{n+1}(x)\mathrm{d}x=0. \tag{5.44}$$

证明 先证必要性. 设 $P(x)$ 是任意次数不超过 n 的多项式，则 $P(x)\omega_{n+1}(x)$ 的次数不超过 $2n+1$. 因此，如果 $x_0,x_1,\cdots,x_n$ 是高斯点，则求积公式(5.41)对于 $P(x)\omega_{n+1}(x)$ 能准确成立，即有

$$\int_a^b\rho(x)P(x)\omega_{n+1}(x)\mathrm{d}x=\sum_{k=0}^{n}\omega_kP(x_k)\omega_{n+1}(x_k),$$

但 $\omega_{n+1}(x_k)=0(k=0,1,\cdots,n)$，故式(5.44)成立.

再证充分性. 对于任意给定的次数不超过 $2n+1$ 的多项式 $f(x)$，用 $\omega_{n+1}(x)$ 除 $f(x)$，记商为 $P(x)$，余式为 $Q(x)$，$P(x)$ 与 $Q(x)$ 都是次数不超过 n 的多项式：

$$f(x)=P(x)\omega_{n+1}(x)+Q(x).$$

而利用式(5.44)，得

$$\int_a^b\rho(x)f(x)\mathrm{d}x=\int_a^b\rho(x)Q(x)\mathrm{d}x \tag{5.45}$$

由于所给求积公式(5.41)是插值型的，它对于 $Q(x)$ 能准确成立：

$$\int_a^b\rho(x)Q(x)\mathrm{d}x=\sum_{k=0}^{n}\omega_kQ(x_k).$$

再注意到 $\omega_{n+1}(x_k)=0(k=0,1,\cdots,n)$，知 $Q(x_k)=f(x_k)(k=0,1,\cdots,n)$，从而由式(5.45)得

$$\int_a^b\rho(x)f(x)\mathrm{d}x=\int_a^b\rho(x)Q(x)\mathrm{d}x=\sum_{k=0}^{n}\omega_kf(x_k)$$

可见求积公式(5.41)对于一切次数不超过 $2n+1$ 的多项式均能准确成立，因此 $x_k(k=0,1,\cdots,n)$ 是高斯点. 定理证毕.

定理表明在 $[a,b]$ 上带权 $\rho(x)$ 的 $n+1$ 次正交多项式的零点就是求积公式(5.41)的高斯点. 有了高斯点 $x_k(k=0,1,\cdots,n)$，再利用式(5.41)，则得到一组关于求积系数 $\omega_k(k=0,1,\cdots,n)$ 的线性代数方程组，由此可得权重系数 $\omega_k(k=0,1,\cdots,n)$. 当然在求出 $x_k(k=0,1,\cdots,n)$ 后，也可由 $x_k(k=0,1,\cdots,n)$ 的插值多项式求出求积系数 $\omega_k(k=0,1,\cdots,n)$.

由上述分析可知，$n+1$ 个节点的插值型求积公式中，高斯求积公式的精度最高.

下面分析高斯求积公式的余项和稳定性.

利用 $f(x)$ 在节点 $x_k(k=0,1,\cdots,n)$ 的 Hermite 插值 $H_{2n+1}(x)$

$$H_{2n+1}(x_k)=f(x_k),H'_{2n+1}(x_k)=f'(x_k)\quad(k=0,1,\cdots,n).$$

从而

$$f(x)=H_{2n+1}(x)+\frac{f^{(2n+2)}(\xi)}{(2n+2)!}\omega_{n+1}^2(x).$$

两端乘以 $\rho(x)$ 并由 a 到 b 积分得到

$$I=\int_a^b\rho(x)f(x)\mathrm{d}x=\int_a^b\rho(x)H_{2n+1}(x)\mathrm{d}x+R_n(f) \tag{5.46}$$

注意到上式右端第一项积分对 $2n+1$ 此多项式准确成立，从而

$$\begin{aligned}R_n(f)&=I-\int_a^b\rho(x)H_{2n+1}(x)\mathrm{d}x=I-\sum_{k=0}^{n}\omega_kH_{2n+1}(x_k)\\&=I-\sum_{k=0}^{n}\omega_kf(x_k)=\int_a^b\rho(x)\frac{f^{(2n+2)}(\xi)}{(2n+2)!}\omega_{n+1}^2(x)\mathrm{d}x\end{aligned}$$

由于 $\rho(x)\omega_{n+1}^2(x)\geqslant 0$，由积分中值定理可得式(5.41)的余项为

$$R_n(f)=\frac{f^{(2n+2)}(\eta)}{(2n+2)!}\int_a^b\rho(x)\omega_{n+1}^2(x)\mathrm{d}x. \tag{5.47}$$

和高阶牛顿—柯特斯公式相比，高斯公式不但精度高，而且是数值稳定的．高斯公式的稳定性之所以能够得到保证，是由于它的求积系数 ω_k 具有非负性．

考察

$$l_k(x)=\prod_{\substack{j=0\\j\neq k}}^{n}\frac{x-x_j}{x_k-x_j},$$

它是 n 次多项式，因而 $l_k^2(x)$ 是 $2n$ 次多项式，故高斯公式(5.41)对于它能准确成立，即有

$$\int_a^b\rho(x)l_k^2(x)\mathrm{d}x=\sum_{i=0}^{n}\omega_il_k^2(x_i).$$

注意到 $l_k(x_i)=\delta_{ki}$，上式右端实际上即等于 ω_k，从而有

$$\omega_k=\int_a^b\rho(x)l_k^2(x)\mathrm{d}x>0$$

由定理 5.2 可知高斯求积公式稳定．

5.5.2 高斯—勒让德求积公式

不失一般性，取 $a=-1$，$b=1$，权函数 $\rho(x)=1$，考察区间$[-1,1]$上的高斯公式

$$\int_{-1}^{1}f(x)\mathrm{d}x\approx\sum_{k=0}^{n}\omega_kf(x_k). \tag{5.48}$$

勒让德多项式(见附录 C.1)是区间$[-1,1]$上的正交多项式，因此，勒让德多项式 $P_{n+1}(x)$ 的零点就是求积公式(5.48)的高斯点．形如式(5.48)的高斯公式称为高斯—勒让德公式．

下面写出 n 较小时的高斯公式．

当 $n=0$ 时，勒让德正交多项式 $P_1(x)=x$，其零点 $x_0=0$．以它为节点构造求积公式

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx \omega_0 f(0),$$

令它对 $f(x)=1$ 准确成立，即可定出 $\omega_0=2$. 从而

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx 2f(0). \tag{5.49}$$

这样构造出的一点高斯—勒让德公式实际上就是中矩形公式.

当 $n=1$ 时，勒让德正交多项式 $P_2(x)=\frac{1}{2}(3x^2-1)$ 的两个零点 $\pm\frac{1}{\sqrt{3}}$ 构造求积公式：

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx \omega_0 f\left(-\frac{1}{\sqrt{3}}\right)+\omega_1 f\left(\frac{1}{\sqrt{3}}\right).$$

令它对 $f(x)=1, x$ 都准确成立，有

$$\begin{cases}\omega_0+\omega_1=2,\\ \omega_0\left(-\dfrac{1}{\sqrt{3}}\right)+\omega_1\left(\dfrac{1}{\sqrt{3}}\right)=0,\end{cases}$$

由此解出 $\omega_0=\omega_1=1$，从而得到两点高斯—勒让德公式

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx f\left(-\frac{1}{\sqrt{3}}\right)+f\left(\frac{1}{\sqrt{3}}\right). \tag{5.50}$$

当 $n=2$ 时，三点高斯—勒让德公式的形式是

$$\int_{-1}^{1} f(x)\mathrm{d}x \approx \frac{5}{9}f\left(-\frac{\sqrt{15}}{5}\right)+\frac{8}{9}f(0)+\frac{5}{9}f\left(\frac{\sqrt{15}}{5}\right). \tag{5.51}$$

当 n 较大时，勒让德多项式的零点只能用数值计算方法求出. 表 5.5 列出了不同 n 对应的高斯点和求积系数.

表 5.5　高斯点与求积系数

n	x_k	ω_k
0	0.0000000	2.0000000
1	±0.5773503	1.0000000
2	±0.7745967 0.0000000	0.5555556 0.8888889
3	±0.8611363 ±0.3399810	0.3478548 0.6521452
4	±0.96061798 ±0.5384693 0.0000000	0.2369269 0.4786287 0.5688889
5	±0.93246951 ±0.66120939 ±0.23861919	0.17132449 0.36076157 0.46791393

对于一般区间$[a,b]$上的积分，可以用变量代换

$$x=\frac{b-a}{2}t+\frac{b+a}{2}, \tag{5.52}$$

将积分区间化为$[-1,1]$，这时

$$\int_a^b f(x)\mathrm{d}x=\frac{b-a}{2}\int_{-1}^1 f\left(\frac{b-a}{2}t+\frac{b+a}{2}\right)\mathrm{d}t,$$

然后对等式右端的积分即可采用高斯—勒让德求积公式.

例 5.14 分别用两点和三点高斯—勒让德公式计算积分

$$I=\int_0^1 \frac{\sin x}{x}\mathrm{d}x.$$

解 先做变量$x=\frac{1}{2}(t+1)$，这时原积分变成了

$$I=\int_0^1 \frac{\sin x}{x}\mathrm{d}x=\int_{-1}^1 \frac{\sin\frac{1}{2}(t+1)}{t+1}\mathrm{d}t.$$

用两点高斯勒让德公式(5.50)可得

$$I\approx\frac{\sin\frac{1}{2}(-0.5773503+1)}{-0.5773503+1}+\frac{\sin\frac{1}{2}(0.5773503+1)}{0.5773503+1}=0.9460411.$$

用三点高斯公式(5.51)可得

$$\begin{aligned}I\approx{}&0.5555556\times\frac{\sin\frac{1}{2}(-0.7745967+1)}{-0.7745967+1}+0.8888889\times\frac{\sin\frac{1}{2}(0+1)}{0+1}\\&+0.5555556\times\frac{\sin\frac{1}{2}(0.7745967+1)}{0.7745967+1}\\={}&0.9460831.\end{aligned}$$

由于I的准确值为0.9460831…，因而三点高斯公式的积分值具有7位有效数字. 而用复化梯形公式(例5.11)，共用了1025个节点上的函数值，才达到了7位有效数字. 用龙贝格公式(例5.12)对区间二分了3次，用了9个节点上的函数值，也得到了7位有效数字. 这充分说明对于同样节点数目的求积公式来说，高斯—勒让德求积公式的计算精度确实高.

例 5.15 将积分区间四等分，用复化两点高斯公式计算积分

$$I=\int_1^3 \frac{1}{x}\mathrm{d}x.$$

解 首先将区间四等分，把积分转化成四个小区间上积分和的形式.

$$I=\int_1^{1.5}\frac{1}{x}\mathrm{d}x+\int_{2.5}^2\frac{1}{x}\mathrm{d}x+\int_2^{2.5}\frac{1}{x}\mathrm{d}x+\int_{2.5}^3\frac{1}{x}\mathrm{d}x$$

记上式右端的四个积分分别为 I_1, I_2, I_3, I_4.

$$I_1 = \int_1^{1.5} \frac{1}{x}\mathrm{d}x = \int_{-1}^{1} \frac{0.5}{2.5+0.5t}\mathrm{d}t$$

$$\approx 0.5 \times \left[\frac{1}{2.5+0.5\times(-1/\sqrt{3})} + \frac{1}{2.5+0.5\times 1/\sqrt{3}}\right]$$

$$\approx 0.405405406$$

类似的计算可得

$$I_2 \approx 0.287671233, I_3 \approx 0.223140496, I_4 \approx 0.182320442,$$

从而

$$I = I_1 + I_2 + I_3 + I_4 \approx 1.098537577.$$

5.5.3 高斯—切比雪夫求积公式

若 $a=-1, b=1$，且取权函数 $\rho(x)=\dfrac{1}{\sqrt{1-x^2}}$，则所建立的高斯公式为

$$\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\mathrm{d}x \approx \sum_{k=0}^{n} \omega_k f(x_k), \tag{5.53}$$

称为高斯—切比雪夫公式. 由于区间$[-1,1]$上关于权函数$\dfrac{1}{\sqrt{1-x^2}}$的正交多项式是切比雪夫多项式，因此求积公式(5.53)的高斯点是 $n+1$ 次切比雪夫多项式的零点，即为

$$x_k = \cos\left(\frac{2k+1}{2n+2}\pi\right) \quad (k = 0,1,\cdots,n)$$

通过计算可知式(5.53)的系数为 $\omega_k=\dfrac{\pi}{n+1}$，使用时将 $n+1$ 个节点公式改为 n 个节点，于是高斯—切比雪夫公式为

$$\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\mathrm{d}x \approx \frac{\pi}{n}\sum_{k=1}^{n} f(x_k), x_k = \cos\frac{(2k-1)\pi}{2n}, \tag{5.54}$$

其余项为

$$R[f] = \frac{2\pi}{2^{2n}(2n)!} f^{(2n)}(\eta) \quad \eta \in [-1,1].$$

例 5.16 用五点高斯—切比雪夫求积公式计算奇异积分

$$I = \int_{-1}^{1} \frac{\mathrm{e}^x}{\sqrt{1-x^2}\mathrm{d}x}$$

解 这里 $f(x)=\mathrm{e}^x$，当 $n=5$ 时，由式(5.54)可得

$$I \approx \frac{\pi}{5}\sum_{k=1}^{5} \exp\left(\cos\frac{(2k-1)\pi}{10}\right) \approx 3.977463.$$

5.5.4 无穷区间上的高斯求积公式

设 $a=0,b=+\infty,\rho(x)=e^{-x}$，利用拉盖尔(Laguerre)多项式可得如下高斯—拉盖尔求积公式

$$\int_0^{+\infty} e^{-x} f(x)\mathrm{d}x \approx \sum_{k=0}^{n} \omega_k f(x_k) \tag{5.55}$$

及

$$\int_0^{+\infty} f(x)\mathrm{d}x \approx \sum_{k=0}^{n} \omega_k e^{x_k} f(x_k), \tag{5.56}$$

其中 $x_k(k=0,1,2,\cdots,n)$ 为 $n+1$ 次拉盖尔多项式的零点.
其余项为

$$R[f]=\frac{[(n+1)!]^2}{(2n+2)!} f^{(2n+2)}(\eta) \quad \eta \in (0,+\infty).$$

高斯—拉盖尔求积公式的高斯点和求积系数见表 5.6.

表 5.6 高斯点与拉盖尔求积系数

n	x_k	ω_k
0	1.0000000000	1.0000000000
1	0.5857864376 3.4142135623	0.85355533905 0.1464466094
2	0.4157745567 2.2942803602 6.2899450829	0.7110930099 0.2785177335 0.0103892565
3	0.3225476896 1.7457611011 4.5366202969 9.3950709123	0.6031541043 0.3574186924 0.0388879085 0.0005392947
4	0.2635603197 1.4134030591 3.5964257710 7.0858100058 12.6408008442	0.5217556105 0.3986668110 0.0759424496 0.0036117586 0.0000233699

设 $a=-\infty,b=+\infty,\rho(x)=e^{-x^2}$，利用埃米特(Hermite)多项式可得如下高斯—埃米特求积公式

$$\int_{-\infty}^{+\infty} e^{-x^2} f(x)\mathrm{d}x \approx \sum_{k=0}^{n} \omega_k f(x_k) \tag{5.57}$$

及

$$\int_{-\infty}^{+\infty} f(x)\mathrm{d}x \approx \sum_{k=0}^{n} \omega_k e^{x_k^2} f(x_k), \tag{5.58}$$

其中 $x_k(k=0,1,2,\cdots,n)$为 $n+1$ 次埃米特多项式的零点.
其余项为

$$R[f]=\frac{(n+1)!\sqrt{\pi}}{2^{n+1}(2n+2)!}f^{(2n+2)}(\eta)\quad \eta\in(-\infty,+\infty).$$

高斯—埃米特求积公式的高斯点和求积系数见表 5.7.

表 5.7　高斯点与埃米特求积系数

n	x_k	ω_k
0	0	1.7724538509
1	+0.7071067811 −0.7071067811	0.8862269254 0.8862269254
2	−1.2247448713 0 +1.2247448713	0.2954089751 1.1816359006 0.2954089751
3	−1.6506801238 −0.5246476232 +0.5246476232 +1.6506801238	0.0813128354 0.8049140900 0.8049140900 0.0813128354
4	−2.0201828704 −0.9585724646 0 +0.9585724646 +2.0201828704	0.0199532420 0.3936193231 0.943087204 0.3936193231 0.0199532420

例 5.17　用高斯—拉盖尔求积公式计算

$$I=\int_0^{+\infty}\mathrm{e}^{-x}\sin x\mathrm{d}x.$$

解　采用四点高斯公式计算可得 $I\approx0.502275$.

例 5.18　用五点高斯—埃米特求积公式计算

$$I=\int_{-\infty}^{+\infty}\mathrm{e}^{-x^2}\cos x\mathrm{d}x.$$

解　采用五点高斯—埃米特公式可得 $I\approx1.380390$.

5.6　数值微分

本节讨论已知 $f(x)$在离散节点处的值 $f(x_k)(k=0,1,2,\cdots,n)$,求 $f(x)$在节点处的导数的近似方法.

由插值理论，可作 $f(x)$ 的 n 次插值多项式 $P_n(x)$，再利用 $P_n(x)$ 的导数来近似代替 $f(x)$ 的导数，即

$$f'(x) \approx P'_n(x). \tag{5.59}$$

此方法称为插值型数值微分法（又称为插值型求导公式）.

必须指出，即使 $f(x)$ 与 $P_n(x)$ 的值相差不多，导数的近似值 $P_n(x)$ 与导数的真值 $f'(x)$ 仍然可能差别很大，因而在使用求导公式(5.59)时应特别注意误差的分析 .

依据插值余项定理，求导公式(5.59)的余项为

$$f'(x) - P'_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega'_{n+1}(x) + \frac{\omega_{n+1}(x)}{(n+1)!}\frac{\mathrm{d}}{\mathrm{d}x}f^{(n+1)}(\xi),$$

其中

$$\omega_{n+1}(x) = \prod_{k=0}^{n}(x - x_k)$$

在这一余项公式中，由于 ξ 是 x 的未知函数，无法对它的第二项 $\frac{\omega_{n+1}(x)}{(n+1)!}\cdot\frac{\mathrm{d}}{\mathrm{d}x}f^{(n+1)}(\xi)$ 作出进一步的说明 . 因此，对于随意给出的点 x，误差 $f'(x)-P'_n(x)$ 是无法预估的 . 但是，如果限定求某个节点 x_k 上的导数值，那么上面的第二项因 $\omega_{n+1}(x_k)=0$ 而变为零，这时有余项公式

$$f'(x_k) - P'_n(x_k) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega'_{n+1}(x_k). \tag{5.60}$$

下面具体写出在等距节点的情况下，节点处导数的计算公式 .

1）两点公式

设已给出两个节点 x_0、x_1 上面的函数值 $f(x_0)$、$f(x_1)$，作线性插值公式

$$P_1(x) = \frac{x - x_1}{x_0 - x_1}f(x_0) + \frac{x - x_0}{x_1 - x_0}f(x_1).$$

对上式两端求导，记 $x_1 - x_0 = h$ 有

$$P'_1(x) = \frac{1}{h}[-f(x_0) + f(x_1)].$$

于是有下列求导公式：

$$P'_1(x_0) = \frac{1}{h}[f(x_1) - f(x_0)], P'_1(x_1) = \frac{1}{h}[f(x_1) - f(x_0)]. \tag{5.61}$$

而利用余项公式(5.60)知，带余项的两点公式是

$$f'(x_0) = \frac{1}{h}[f(x_1) - f(x_0)] - \frac{h}{2}f''(\xi),$$

$$f'(x_1) = \frac{1}{h}[f(x_1) - f(x_0)] + \frac{h}{2}f''(\xi).$$

2)三点公式

设已给出三个节点 $x_0, x_1=x_0+h, x_2=x_0+2h$ 上的函数值，作二次插值

$$P_2(x)=\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}f(x_0)+\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}f(x_1)$$

$$+\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}f(x_2).$$

令 $x=x_0+th$，上式可表为

$$P_2(x_0+th)=\frac{1}{2}(t-1)(t-2)f(x_0)-t(t-2)f(x_1)+\frac{1}{2}t(t-1)f(x_2),$$

两端对 t 求导，有

$$P_2'(x_0+th)=\frac{1}{2h}[(2t-3)f(x_0)-(4t-4)f(x_1)+(2t-1)f(x_2)]. \tag{5.62}$$

这里撇号表示对变量 x 求导数．上式分别取 $t=0,1,2$，得到三种三点公式：

$$P_2'(x_0)=\frac{1}{2h}[-3f(x_0)+4f(x_1)-2f(x_2)],$$

$$P_2'(x_1)=\frac{1}{2h}[-f(x_0)+f(x_2)],$$

$$P_2'(x_2)=\frac{1}{2h}[f(x_0)-4f(x_1)+3f(x_2)].$$

而带余项的三点求导公式如下：

$$f'(x_0)=\frac{1}{2h}[-3f(x_0)+4f(x_1)-f(x_2)]+\frac{h^2}{3}f'''(\xi),$$

$$f'(x_1)=\frac{1}{2h}[-f(x_0)+f(x_2)]-\frac{h^2}{6}f'''(\xi), \tag{5.63}$$

$$f'(x_2)=\frac{1}{2h}[f(x_0)-4f(x_1)+3f(x_2)]+\frac{h^2}{3}f'''(\xi).$$

其中的式(5.63)是中点公式．

用插值多项式 $P_n(x)$ 作为 $f(x)$ 的近似函数，还可以建立高阶数值微分公式：

$$f^{(k)}(x)\approx P_n^{(k)}(x)\quad(k=1,2,\cdots).$$

例如，将式(5.62)再对 t 求导一次，有

$$P_2''(x_0+th)=\frac{1}{h^2}[f(x_0)-2f(x_1)+f(x_2)],$$

于是有

$$P_2''(x_1)=\frac{1}{h^2}[f(x_1-h)-2f(x_1)+f(x_1+h)]. \tag{5.64}$$

而带余项的二阶三点公式如下：

$$f''(x_1)=\frac{1}{h^2}[f(x_1-h)-2f(x_1)+f(x_1+h)]-\frac{h^2}{12}f^{(4)}(\xi). \tag{5.65}$$

除了上述插值型数值微分方法外，还可以利用直接差商法求得函数在某节点处的导数的近似值．该种方法将在微分方程数值解中讲到．

例 5.19 给定函数 $f(x)=e^x$ 的数据见表 5.8，试利用两点、三点微分公式计算 $x=2.7$ 处的一阶、二阶导数值．

表 5.8 函数值

x	2.5	2.6	2.7	2.8	2.9
$f(x)$	12.1825	13.4637	14.8797	16.4446	18.1741

解 由于题目中没有指明用哪些节点来计算 $f'(2.7)$ 和 $f''(2.7)$，因此，取不同的节点，会得到不同的计算结果．

在两点公式中，取 $x_0=2.6, x_1=2.7, h=0.1$，则

$$f'(2.7)\approx\frac{1}{0.1}[f(2.7)-f(2.6)]=14.1600.$$

若取 $x_0=2.7, x_1=2.8, h=0.1$，则

$$f'(2.7)\approx\frac{1}{0.1}[f(2.8)-f(2.7)]=15.6490.$$

在三点公式中，取 $x_0=2.6, x_1=2.7, x_2=2.8, h=0.1$，则

$$f'(2.7)\approx\frac{1}{2\times0.1}[f(2.8)-f(2.6)]=14.9045,$$

$$f''(2.7)\approx\frac{1}{0.1^2}[f(2.8)-2f(2.7)+f(2.6)]=14.8900.$$

另外，若在两点公式中取 $x_0=2.5, x_1=2.7, h=0.2$，则

$$f'(2.7)\approx\frac{1}{0.2}[f(2.7)-f(2.5)]=13.4860.$$

若取 $x_0=2.7, x_1=2.9, h=0.2$，则

$$f'(2.7)\approx\frac{1}{0.2}[f(2.9)-f(2.7)]=16.4720.$$

在三点公式中，取 $x_0=2.5, x_1=2.7, x_2=2.9, h=0.2$，则

$$f'(2.7)\approx\frac{1}{2\times0.2}[f(2.9)-f(2.5)]=14.9790,$$

$$f''(2.7)\approx\frac{1}{0.2^2}[f(2.9)-2f(2.7)+f(2.5)]=14.9300.$$

将上述结果与 $f'(2.7)$ 和 $f''(2.7)$ 的准确值 14.879730 对比可知，步长越小，一般误差也越小．

习　题

1. 确定下列求积公式中的待定参数，使其代数精度尽量高，并指出其代数精度．

(1)$\int_0^{2h} f(x)\mathrm{d}x \approx A_0 f(0) + A_1 f(h) + A_2 f(2h)$；

(2)$\int_{-1}^{1} f(x)\mathrm{d}x \approx A[f(-1) + 2f(x_1) + 3f(x_2)]$.

2. 分别利用梯形公式、辛普森公式和柯特斯公式计算积分 $\int_0^1 x^2\mathrm{d}x$．

3. 分别利用复化梯形公式和复化辛普森公式计算积分 $\int_0^1 \frac{1}{1+x}\mathrm{d}x$．

4. 按下列指定公式，求积分 $\int_1^2 \frac{1}{x}\mathrm{d}x$ 的近似值，并与 ln2 比较．

(1)龙贝格公式(二分 3 次)；

(2)三点高斯—勒让德公式；

(3)$n=4$ 时的复化两点高斯—勒让德公式．

5. 推导下列三种矩形求积公式．

(1)$\int_a^b f(x)\mathrm{d}x = (b-a)f(a) + \frac{1}{2}f'(\xi)(b-a)^2$；

(2)$\int_a^b f(x)\mathrm{d}x = (b-a)f(b) - \frac{1}{2}f'(\zeta)(b-a)^2$；

(3)$\int_a^b f(x)\mathrm{d}x = (b-a)f\left(\frac{a+b}{2}\right) + \frac{1}{24}f''(\eta)(b-a)^3$.

6. 用泰勒级数展开的方法计算积分 $\int_0^1 \mathrm{e}^{-x^2}\mathrm{d}x$，要求计算精度为 10^{-4}．

7. 分别利用高斯—拉盖尔求积公式或高斯—埃尔米特求积公式计算下列积分．

(1)$\int_0^{\infty} \frac{x\mathrm{e}^{-x}}{x+2}\mathrm{d}x \quad (n=4)$；

(2)$\int_{-\infty}^{+\infty} \frac{x+1}{x+2}\mathrm{e}^{-x^2}\mathrm{d}x$.

8. 已知函数 $f(x)=\frac{1}{(1+x)^2}$ 的值由表 5.9 给出．

表 5.9　函数值

x	1.0	1.1	1.2
$f(x)$	0.2500	0.2268	0.2056

用三点公式求函数在 $x=1.0, 1.1$ 和 1.2 处的导数值，并估计误差．

6 解线性方程组的直接法

6.1 引言

在自然科学和工程技术中，很多问题的求解常常归结为解线性代数方程组．例如第 4 章中用最小二乘法求实验数据的曲线拟合问题，用差分法或者有限元方法解常微分方程、偏微分方程边值问题，第 5 章例 5.9 用数值积分解积分方程等都导致求解线性代数方程组．

对于方程组的求解问题，大家似乎都比较熟悉．对于一个有唯一解的线性代数方程组，可利用 Cramer 法则，或利用初等行变换将方程组转化成同解的上三角方程组来求解．下面将由具体的例子来分析在求解方程组的时候会出现一些亟待解决的问题．

例 6.1 下面考虑 n 阶方程组 $\boldsymbol{Ax}=\boldsymbol{b}$

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2, \\ \cdots\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n. \end{cases} \tag{6.1}$$

这里系数矩阵 $\boldsymbol{A}=(a_{ij})_{n\times n}$，$\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$，$\boldsymbol{b}=(b_1,b_2,\cdots,b_n)^{\mathrm{T}}$．

当时 $\det \boldsymbol{A}=D\neq 0$ 时，由线性代数方程组中的 Cramer 法则可知，方程组的解存在唯一，且

$$x_i = \frac{D_i}{D} \quad (i = 1,2,\cdots,n).$$

D_i 为系数矩阵 $\boldsymbol{A}$ 的第 i 列元素以 $\boldsymbol{b}$ 代替的矩阵的行列式的值．Cramer 法则在线性代数建立方程组解的理论基础中功不可没，但是在实际计算中，其计算量是惊人的．

该例表明，尽管知道如何求解线性代数方程组，但是还得考虑计算量和求解效率．在上一个例子中，如果选用高斯消元法，只用几秒钟而已．

例 6.2 求解两个方程组

$$\begin{cases} x_1 + x_2 = 2, \\ x_1 + 1.0001x_2 = 2, \end{cases} \tag{6.2}$$

$$\begin{cases} x_1 + x_2 = 2, \\ x_1 + 1.0001x_2 = 2.0001. \end{cases} \tag{6.3}$$

式(6.2)的精确解是 $(x_1,x_2)^{\mathrm{T}}=(2,0)^{\mathrm{T}}$，而式(6.3)的精确解是 $(x_1,x_2)^{\mathrm{T}}=(1,1)^{\mathrm{T}}$．两个方程组的系数矩阵相同，右端项也只是第二个数据做了微小的变化，但是方程组的解却发生了很大的变化．第 4 章中求解曲线拟合的法方程经常就会遇到这种类型的方程组．这种方程组称为

病态方程组，系数矩阵 $\boldsymbol{A}$ 称为病态矩阵．

例 6.3 第 1 章中单相渗流的连续性方程是一个三维抛物型方程，下面考虑一个一维的抛物型方程的初边值问题．

$$\begin{cases} \dfrac{\partial u}{\partial t} = a\dfrac{\partial^2 u}{\partial x^2} + f(x) \quad (0 < t \leqslant T), \\ u(x,0) = \varphi(x), \\ u(0,t) = u(l,t) = 0. \end{cases} \tag{6.4}$$

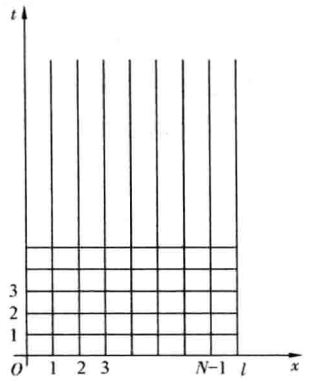

图 6.1 网格示意图

对于给定的 $f(x)$ 和 $\varphi(x)$，可以采用分离变量法求出方程的准确解．而现在考虑的是利用有限差分法求方程组的差分解．

取空间步长 $h=l/N$，时间步长 $\tau=T/M$，其中 N、M 都是正整数．用两族平行直线 $x=x_j=jh(j=0,1,2,\cdots,N)$ 和 $t=t_k=k\tau(k=0,1,2.\cdots,M)$ 将矩形区域 $\{0\leqslant x\leqslant l, 0\leqslant t\leqslant T\}$ 分割成矩形网格(图 6.1)，网格节点为 (x_j,t_k)．

有限差分法就是求出函数 u 在网格节点处函数的近似值．由于矩形区域的边界上函数值均已知，所以只需要求出函数网格内点(位于矩形区域内部的节点)上的近似值．若以 u_j^k 表示 u 在节点 (x_j,t_k) 处的近似值，利用差商代替微商，可以得到一个非常著名的两层隐式绝对稳定的六点对称有限差分格式(也称为 Crank—Nicolson 格式)．

$$\frac{u_j^{k+1}-u_j^k}{\tau} = \frac{a}{2}\left[\frac{u_{j+1}^{k+1}-2u_j^{k+1}+u_{j-1}^{k+1}}{h^2} + \frac{u_{j+1}^k-2u_j^k+u_{j-1}^k}{h^2}\right] + f_j, \tag{6.5}$$

$$u_j^0 = \varphi_j = \varphi(x_j), u_0^k = u_N^k = 0.$$

其中 $j=1,2,\cdots,N-1, k=1,2,\cdots,M-1$．记 $r=a\tau/h^2$，称为网格比．将式(6.5)改写成

$$-\frac{r}{2}u_{j+1}^{k+1} + (1+r)u_j^{k+1} - \frac{r}{2}u_{j-1}^{k+1} = \frac{r}{2}u_{j+1}^k + (1-r)u_j^k + \frac{r}{2}u_{j-1}^{k+1} + \tau f_j \ . \tag{6.6}$$

利用 u_j^0 和边界条件便可逐层求得 u_j^k．差分格式(6.6)就是一个线性代数方程组，其系数矩阵为

$$\begin{bmatrix} 1+r & -\frac{r}{2} & 0 & 0 & 0 \\ -\frac{r}{2} & 1+r & -\frac{r}{2} & 0 & 0 \\ \ddots & \ddots & \ddots & 0 & 0 \\ 0 & -\frac{r}{2} & 1+r & -\frac{r}{2} & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & -\frac{r}{2} & 1+r & -\frac{r}{2} \\ 0 & 0 & 0 & -\frac{r}{2} & 1+r \end{bmatrix}.$$

这个系数矩阵有自己的特点，称为三对角矩阵，而且严格对角占优，因此方程组式(6.6)是一个严格对角占优的三对角方程组，这种方程组有没有特殊的解法？

由上述例子可以看出，方程组的求解并不是一个简单问题．由于这些方程组的系数矩阵的特点不一样，有的是低阶稠密矩阵(例如，阶数大约为≤150)，有的是大型稀疏矩阵(即矩阵阶数高且零元素较多，如三对角矩阵)，有的是对称正定矩阵，所以求解方法也有所差异．如何利用计算机来快速、有效地求解线性代数方程组是数值线性代数研究的核心问题．

关于线性方程组的数值解法一般有两类：

(1)直接法，就是经过有限步算术运算，可求得方程组精确解的方法(若计算过程中没有舍入误差)．但实际计算中由于舍入误差的存在和影响，这种方法也只能求得线性方程组的近似解．本章将阐述这类算法中最基本的高斯消去法及其高斯列主元消去法．这类方法是解低阶稠密矩阵方程组的有效方法，近十几年来直接法在求解具有较大型稀疏矩阵方程组方面取得了较大进展．

(2)迭代法，就是用某种极限过程去逐步逼近线性方程组精确解的方法．迭代法具有需要计算机的存贮单元较少、程序设计简单、原始系数矩阵在计算过程中始终不变等优点，但存在收敛性及收敛速度问题．迭代法是解大型稀疏矩阵方程组(尤其是由微分方程离散后得到的大型方程组)的重要方法．

6.2 高斯消去法

本节介绍高斯消去法(逐次消去法)以及消去法和矩阵的三角分解之间的关系．虽然高斯消去法是一个古老的求解线性方程组的方法(早在公元前250年我国就掌握了解三元一次联立方程组的方法)，但由它改进、变形得到的主元素消去法、三角分解法仍然是目前计算机上常用的有效方法．

6.2.1 高斯消去法的内容

形如下面两种形式的方程组的求解是非常容易的．

(1)下三角方程组

$$\begin{cases} l_{11}x_1 = b_1, \\ l_{21}x_1 + l_{22}x_2 = b_2, \\ \cdots\cdots \\ l_{n1}x_1 + l_{n2}x_2 + \cdots + l_{nn}x_n = b_n. \end{cases} \tag{6.7}$$

设 $l_{ii} \neq 0$，按照方程组的顺序，由第一个方程解出 x_1，将 x_1 代入第二个方程解出 x_2，将 x_1、x_2 代入第三个方程解出 x_3，依此类推，最后解出 x_n．一般的求解公式为

$$x_i = \left(b_i - \sum_{j=1}^{i-1} l_{ij}x_j\right)/l_{ii} \quad (i = 1,2,\cdots,n).$$

(2)上三角方程组

$$\begin{cases} u_{11}x_1 + u_{12}x_2 + \cdots + u_{1n}x_n = b_1, \\ u_{22}x_2 + \cdots + u_{2n}x_n = b_2, \\ \cdots\cdots \\ u_{mn}x_n = b_n. \end{cases} \tag{6.8}$$

与解下三角方程组的次序相反，首先从第 n 个方程解出 x_n，将 x_n 代入第 $n-1$ 个方程解出 x_{n-1}，然后将 x_n、x_{n-1} 代入第 $n-2$ 个方程解出 x_{n-2}，依此类推，最后解出 x_1. 一般的求解公式为

$$x_i = (b_i - \sum_{j=i+1}^{n} u_{ij}x_j)/u_{ii} \quad (i = n, n-1, \cdots, 1).$$

一般的方程组不是上面两种形式，该如何处理呢？基本思路是把一个一般的方程组式(6.1)借助于矩阵初等行变换转化成上述方程组中的一个.

所谓矩阵初等行变换就是指：

(1)交换一个矩阵两行(对应交换一个方程组的两个方程)；

(2)矩阵的某行乘以一个非零的数(对应一个方程的两边乘以一个非零数)；

(3)矩阵的某行乘以一个数加到另一行上(对应方程组的某个方程乘以一个数加到另一个方程上).

高斯消去法就是通过对方程组做初等行变换，把一个一般的方程组转化成上三角形方程组.

首先举一个简单的例子来说明消去法的基本思想.

例 6.4 用消去法解方程组

$$\begin{cases} x_1 + x_2 + x_3 = 6, \\ 4x_2 - x_3 = 5, \\ 2x_1 - 2x_2 + x_3 = 1. \end{cases}$$

第一步，将第一个方程乘上 -2 加到第三个方程上去，消去第三个方程中的未知数 x_1，得到

$$-4x_2 - x_3 = -11. \tag{6.9}$$

第二步，将第二个方程加到式(6.9)上去，消去式(6.9)中的未知数 x_2，得到与原方程组等价的三角形方程组

$$\begin{cases} x_1 + x_2 + x_3 = 6, \\ 4x_2 - x_3 = 5, \\ -2x_3 = -6. \end{cases} \tag{6.10}$$

显然方程组式(6.10)是上三角形方程组，求解是容易的，解为 $(x_1, x_2, x_3)^{\mathrm{T}} = (1, 2, 3)^{\mathrm{T}}$. 上述过程相当于

$$(\boldsymbol{A}\mid\boldsymbol{b})=\left[\begin{array}{ccc:c}1&1&1&6\\0&4&-1&5\\2&-2&1&1\end{array}\right]\to\left[\begin{array}{ccc:c}1&1&1&6\\0&4&-1&5\\0&-4&-1&-11\end{array}\right]\to\left[\begin{array}{ccc:c}1&1&1&6\\0&4&-1&5\\0&0&-2&-6\end{array}\right],$$

$$(-2)\times r_1+r_3\to r_3,\ r_2+r_3\to r_3.$$

由此看出，用消去法解方程组的基本思想是用逐次消去未知数的方法把原来方程组 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$ 化为与其等价的三角形方程组，而求解三角形方程组就容易了．换句话说，上述过程就是用行的初等变换将原方程组系数矩阵化为简单形式，从而将求解原方程组的问题转化为求解简单方程组的问题．

下面讨论一般的解 n 阶方程组的高斯消去法(高斯消元法)．

将式(6.1)记为 $\boldsymbol{A}^{(1)}\boldsymbol{x}=\boldsymbol{b}^{(1)}$，其中 $\boldsymbol{A}^{(1)}=(a_{ij}^{(1)})=(a_{ij})$，$\boldsymbol{b}^{(1)}=\boldsymbol{b}$.

6.2.1.1 消元过程

(1)第一次消元．设 $a_{11}^{(1)}\neq 0$，首先对行计算 $m_{i1}=a_{i1}^{(1)}/a_{11}^{(1)}$ $(i=2,3,\cdots,n)$．用 $-m_{i1}$ 乘式(6.1)的第一个方程，加到第 i 个$(i=2,3,\cdots,n)$方程上，消去式(6.1)的第二个方程直到第 n 个方程中的未知数 x_1，得到与式(6.1)等价的方程组

$$\begin{bmatrix}a_{11}^{(1)}&a_{12}^{(1)}&\cdots&a_{1n}^{(1)}\\0&a_{22}^{(2)}&\cdots&a_{2n}^{(2)}\\\vdots&\vdots&&\vdots\\0&a_{n2}^{(2)}&\cdots&a_{nn}^{(2)}\end{bmatrix}\begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix}=\begin{bmatrix}b_1^{(1)}\\b_2^{(2)}\\\vdots\\b_n^{(2)}\end{bmatrix}.\tag{6.11}$$

简记为 $\boldsymbol{A}^{(2)}\boldsymbol{x}=\boldsymbol{b}^{(2)}$，其中

$$a_{ij}^{(2)}=a_{ij}^{(1)}-m_{i1}a_{1j}^{(1)}\quad(i,j=2,3,\cdots,n),$$

$$b_i^{(2)}=b_i^{(1)}-m_{i1}b_1^{(1)}\quad(i=2,3,\cdots,n).$$

(2)一般第 k 次消元$(1\leqslant k\leqslant n-1)$.

设第 $k-1$ 步计算已经完成，即已计算好与式(6.1)等价的方程组

$$\boldsymbol{A}^{(k)}\boldsymbol{x}=\boldsymbol{b}^{(k)},\tag{6.12}$$

其中 $\boldsymbol{A}^{(k)}$ 具有形式

$$\boldsymbol{A}^{(k)}=\begin{bmatrix}a_{11}^{(1)}&a_{12}^{(1)}&\cdots&\cdots&\cdots&a_{1n}^{(1)}\\&a_{22}^{(2)}&\cdots&\cdots&\cdots&a_{2n}^{(2)}\\&&\ddots&&&\vdots\\&&&a_{kk}^{(k)}&\cdots&a_{kn}^{(k)}\\&&&\cdots&\cdots&\cdots\\&&&a_{nk}^{(k)}&\cdots&a_{nn}^{(k)}\end{bmatrix}.$$

设 $a_{kk}^{(k)}\neq 0$，计算乘数 $m_{ik}=a_{ik}^{(k)}/a_{kk}^{(k)}(i=k+1,\cdots,n)$，用 $-m_{ik}$ 乘式(6.12)的第 k 个方程加到第 $i(i=k+1,\cdots,n)$ 个方程上，消去第 $k+1$ 个方程直到第 n 个方程的未知数 x_k 得到与式(6.1)等价的方程组 $\boldsymbol{A}^{(k+1)}\boldsymbol{x}=\boldsymbol{b}^{(k+1)}$.

$\boldsymbol{A}^{(k+1)}$ 和 $\boldsymbol{b}^{(k+1)}$ 的元素的计算公式为

$$a_{ij}^{(k+1)}=a_{ij}^{(k)}-m_{ik}a_{kj}^{(k)}\quad(i,j=k+1,\cdots,n),$$

$$b_i^{(k+1)}=a_i^{(k)}-m_{ik}b_k^{(k)}\quad(i=k+1,\cdots,n).$$

显然 $\boldsymbol{A}^{(k+1)}$ 的第 1 行到第 k 行与 $\boldsymbol{A}^{(k)}$ 相同.

(3)继续这一过程，且设 $a_{ii}^{(i)}\neq 0(i=2,3,\cdots,n-1)$（$a_{ii}^{(i)}$ 称为约化的主元素），直到完成第 $n-1$ 次消元. 最后得到与原方程组等价的三角形方程组 $\boldsymbol{A}^{(n)}\boldsymbol{x}=\boldsymbol{b}^{(n)}$，即

$$\begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ & & \ddots & \vdots \\ & & & a_{nn}^{(n)} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}=\begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_n^{(n)} \end{bmatrix}. \tag{6.13}$$

由式(6.1)约化为式(6.13)的过程称为消元过程.

6.2.1.2 回代过程

如果 $\boldsymbol{A}\in\boldsymbol{R}^{n\times n}$ 是可逆矩阵且 $a_{ii}^{(i)}\neq 0(i=2,3,\cdots,n)$，求解上三角形方程组式(6.13)，得到求解公式

$$\begin{cases} x_n=b_n^{(n)}/a_{nn}^{(n)}, \\ x_k=\left(b_k^{(k)}-\sum\limits_{j=k+1}^{n}a_{kj}^{(k)}x_j\right)/a_{kk}^{(k)}\quad(k=n-1,n-2,\cdots,1). \end{cases} \tag{6.14}$$

式(6.13)的求解过程式(6.14)称为回代过程.

如果 $a_{11}^{(1)}=0$，由于 $\boldsymbol{A}$ 为非奇异矩阵，所以 $\boldsymbol{A}$ 的第一列一定有元素不等于零，例如 $a_{i_1 1}\neq 0$，于是可交换两行元素($r_1\leftrightarrow r_{i_1}$)，将 $a_{i_1 1}$ 调到(1,1)位置，然后进行消元计算，这时 $\boldsymbol{A}^{(2)}$ 右下角矩阵为 $n-1$ 阶非奇异矩阵. 继续这一过程，高斯消去法照样可进行计算.

6.2.2 高斯列主元素消去法

由高斯消去法可知，在消元过程中可能出现 $a_{kk}^{(k)}=0$ 的情况，这时消去法将无法进行；即使主元素 $a_{kk}^{(k)}\neq 0$ 且很小时，用其作除数，也会导致其他元素数量级的严重增长和舍入误差的扩散，最后也使得计算解不可靠.

例 6.5 求解方程组

$$\begin{bmatrix} 0.001 & 2.000 & 3.000 \\ -1.000 & 3.712 & 4.623 \\ -2.000 & 1.072 & 5.643 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}=\begin{bmatrix} 1.000 \\ 2.000 \\ 3.000 \end{bmatrix}.$$

用 4 位浮点数进行计算．精确解舍入到 4 位有效数字为

$$\boldsymbol{x}^* = (-0.4904, -0.05104, 0.3675)^{\mathrm{T}}.$$

解 方法 1：用高斯消去法求解．

$$(\boldsymbol{A} \mid \boldsymbol{b}) = \left[\begin{array}{ccc:c} 0.001 & 2.000 & 3.000 & 1.000 \\ -1.000 & 3.712 & 4.623 & 2.000 \\ -2.000 & 1.072 & 5.643 & 3.000 \end{array}\right] \begin{array}{l} \\ m_{21} = -1.000/0.001 = -1000 \\ m_{31} = -2.000/0.001 = -2000 \end{array}$$

$$\rightarrow \left[\begin{array}{ccc:c} 0.001 & 2.000 & 3.000 & 1.000 \\ 0 & 2004 & 3005 & 1002 \\ 0 & 4001 & 6006 & 2003 \end{array}\right] m_{32} = 4001/2004 = 1.997$$

$$\rightarrow \left[\begin{array}{ccc:c} 0.001 & 2.000 & 3.000 & 1.000 \\ 0 & 2004 & 3005 & 1002 \\ 0 & 0 & 5.000 & 2.000 \end{array}\right],$$

得计算解为

$$\bar{\boldsymbol{x}} = (-0.400, -0.09980, 0.4000)^{\mathrm{T}}.$$

显然计算解 $\bar{x}$ 是一个很坏的结果，不能作为方程组的近似解．其原因是在消元计算时用了小主元 0.001，使得约化后的方程组元素数量级大大增长，经消元后得到的三角形方程组就不准确了．

方法 2：交换行，避免绝对值小的主元作除数．

$$\underset{r_1 \leftrightarrow r_3}{(\boldsymbol{A} \mid \boldsymbol{b})} = \left[\begin{array}{ccc:c} -2.00 & 1.072 & 5.643 & 3.000 \\ -1.000 & 3.712 & 4.623 & 2.000 \\ 0.001 & 2.000 & 3.000 & 1.000 \end{array}\right] \begin{array}{l} \\ m_{21} = 0.5000 \\ m_{31} = -0.500 \end{array}$$

$$\rightarrow \left[\begin{array}{ccc:c} -2.000 & 1.072 & 5.643 & 3.000 \\ 0 & 3.176 & 1.801 & 0.5000 \\ 0 & 2.001 & 3.003 & 1.002 \end{array}\right] m_{32} = 0.6300$$

$$\rightarrow \left[\begin{array}{ccc:c} -2.000 & 1.072 & 5.643 & 3.000 \\ 0 & 3.176 & 1.801 & 0.5000 \\ 0 & 0 & 1.868 & 0.6870 \end{array}\right],$$

得计算解为

$$\boldsymbol{x} = (-0.4900, -0.05113, 0.3678)^{\mathrm{T}} \approx \boldsymbol{x}^*.$$

由此例子可知，在采用高斯消去法解方程组时，小主元可能产生麻烦，故应避免采用绝对值小的主元素 $a_{kk}^{(k)}$．对一般矩阵来说，最好每步选取系数矩阵（或消元后的低阶矩阵）中绝对值最大的元素作为主元素，以使高斯消去法具有较好的数值稳定性．

下面介绍高斯列主元素消去法．本节中总假定式(6.1)的 $\boldsymbol{A} \in \boldsymbol{R}^{n \times n}$ 非奇异．

设方程组式(6.1)的增广矩阵为

$$\boldsymbol{B}=(\boldsymbol{A}\mid\boldsymbol{b})=\left[\begin{array}{cccc:c}a_{11}&a_{12}&\cdots&a_{1n}&b_1\\a_{21}&a_{22}&\cdots&a_{2n}&b_2\\\vdots&\vdots&&\vdots&\vdots\\a_{n1}&a_{n2}&\cdots&a_{nn}&b_n\end{array}\right]$$

首先在 $\boldsymbol{A}$ 的第一列中选取绝对值最大的元素作为主元素，例如

$$|a_{i_1 1}|=\max_{1\leqslant i\leqslant n}|a_{i1}|\neq 0.$$

然后交换矩阵 $\boldsymbol{B}$ 的第 1 行与第 i_1 行，经第一次消元计算得到

$$(\boldsymbol{A}\mid\boldsymbol{b})\rightarrow(\boldsymbol{A}^{(2)}\mid\boldsymbol{b}^{(2)}).$$

重复上述过程，设已完成第 $k-1$ 步的选主元素，交换两行及消元计算，$(\boldsymbol{A}|\boldsymbol{b})$约化为

$$(\boldsymbol{A}^{(k)}\mid\boldsymbol{b}^{(k)})=\left[\begin{array}{ccccc:c}a_{11}^{(1)}&a_{12}^{(1)}&&&\cdots&a_{1n}^{(1)}&b_1^{(1)}\\&a_{22}^{(2)}&&&\cdots&a_{2n}^{(2)}&b_2^{(2)}\\&&\ddots&&&&\\&&&a_{kk}^{(k)}&\cdots&a_{kn}^{(k)}&b_k^{(k)}\\&&&&&\vdots&\vdots\\&&&a_{nk}^{(k)}&\cdots&a_{nn}^{(k)}&b_n^{(k)}\end{array}\right],$$

且 $\boldsymbol{A}^{(k)}x=\boldsymbol{b}^{(k)}$ 与 $\boldsymbol{A}x=\boldsymbol{b}$ 等价，第 k 步选主元素(在 $\boldsymbol{A}^{(k)}$ 右下角方阵的第 1 列内选取)，即确定i_k 使

$$|a_{i_k,k}^{(k)}|=\max_{k\leqslant i\leqslant n}|a_{ik}^{(k)}|\neq 0$$

交换$(\boldsymbol{A}^{(k)}|\boldsymbol{b}^{(k)})$第 k 行与第 i_k 行的元素，再进行消元计算，最后将方程组化为

$$\left[\begin{array}{cccc}a_{11}^{(1)}&a_{12}^{(1)}&\cdots&a_{1n}^{(1)}\\&a_{22}^{(2)}&\cdots&a_{2n}^{(2)}\\&&\ddots&\vdots\\&&&a_{nn}^{(n)}\end{array}\right]\left[\begin{array}{c}x_1\\x_2\\\vdots\\x_n\end{array}\right]=\left[\begin{array}{c}b_1^{(1)}\\b_2^{(2)}\\\vdots\\b_n^{(n)}\end{array}\right],\tag{6.15}$$

然后再进行回代计算．注意式(6.15)中的元素和式(6.13)中的元素只是形式相同，数值上不一定相等．

6.2.3 矩阵的三角分解

下面借助矩阵理论进一步对高斯消去法作些分析，从而建立高斯消去法与矩阵因式分解的关系．

设式(6.1)中 $\boldsymbol{A}$ 的各顺序主子式均不为零(可以保证高斯消去法中约化的所有主元素

$a_{ii}^{(i)} \neq 0$). 由于对 $\boldsymbol{A}$ 施行初等行变换相当于用相应的初等矩阵左乘 $\boldsymbol{A}$，于是对式(6.1)施行第一次消元后化为式(6.11)，这时 $\boldsymbol{A}^{(1)}$ 化为 $\boldsymbol{A}^{(2)}$，$\boldsymbol{b}^{(1)}$ 化为 $\boldsymbol{b}^{(2)}$，即

$$\boldsymbol{L}_1\boldsymbol{A}^{(1)} = \boldsymbol{A}^{(2)}, \boldsymbol{L}_1\boldsymbol{b}^{(1)} = \boldsymbol{b}^{(2)},$$

其中

$$\boldsymbol{L}_1 = \begin{bmatrix} 1 & & & & \\ -m_{21} & 1 & & & \\ -m_{31} & & 1 & & \\ \vdots & & & \ddots & \\ -m_{n1} & & & & 1 \end{bmatrix}$$

一般第 k 步消元，$\boldsymbol{A}^{(k)}$ 化为 $\boldsymbol{A}^{(k+1)}$，$\boldsymbol{b}^{(k)}$ 化为 $\boldsymbol{b}^{(k+1)}$，相当于

$$\boldsymbol{L}_k\boldsymbol{A}^{(k)} = \boldsymbol{A}^{(k+1)}, \boldsymbol{L}_k\boldsymbol{b}^{(k)} = \boldsymbol{b}^{(k+1)},$$

其中

$$\boldsymbol{L}_k = \begin{bmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & -m_{k+1,k} & 1 & & \\ & & \vdots & & \ddots & \\ & & -m_{nk} & & & 1 \end{bmatrix}.$$

重复这过程，最后得到

$$\boldsymbol{L}_{n-1}\cdots\boldsymbol{L}_2\boldsymbol{L}_1\boldsymbol{A}^{(1)} = \boldsymbol{A}^{(n)},$$

$$\boldsymbol{L}_{n-1}\cdots\boldsymbol{L}_2\boldsymbol{L}_1\boldsymbol{b}^{(1)} = \boldsymbol{b}^{(n)}.$$

将上三角矩阵 $\boldsymbol{A}^{(n)}$，记为 $\boldsymbol{U}$，由上式可得 $\boldsymbol{A}=\boldsymbol{L}_1^{-1}\boldsymbol{L}_2^{-1}\cdots\boldsymbol{L}_{n-1}^{-1}\boldsymbol{U}=\boldsymbol{L}\boldsymbol{U}$，其中

$$\boldsymbol{L} = \boldsymbol{L}_1^{-1}\boldsymbol{L}_2^{-1}\cdots\boldsymbol{L}_{n-1}^{-1} = \begin{bmatrix} 1 & & & & \\ m_{21} & 1 & & & \\ m_{31} & m_{32} & 1 & & \\ \vdots & \vdots & & \ddots & \\ m_{n1} & m_{n2} & \cdots & m_{n,n-1} & 1 \end{bmatrix}$$

为单位下三角矩阵.

这就是说，高斯消去法实质上产生了一个将 $\boldsymbol{A}$ 分解为两个三角形矩阵相乘的因式分解，于是得到如下重要定理，它在解方程组的直接法中起着重要作用.

定理 6.1 (矩阵的 $\boldsymbol{LU}$ 分解)设 $\boldsymbol{A}$ 为 n 阶矩阵，如果 $\boldsymbol{A}$ 的顺序主子式 $D_i \neq 0 (i=1,2,\cdots,$

$n-1$),则 $\boldsymbol{A}$ 可分解为一个单位下三角矩阵 $\boldsymbol{L}$ 和一个上三角矩阵 $\boldsymbol{U}$ 的乘积,且这种分解是唯一的.

矩阵的这种 $\boldsymbol{LU}$ 分解又称为杜利脱尔(Doolittle)分解. 从矩阵的 $\boldsymbol{LU}$ 分解出发还可以得到一些变形情况.

定理 6.2 如果矩阵 $\boldsymbol{A}$ 的各阶顺序主子式不为零,则有:

(1)$\boldsymbol{A}$ 有唯一的三角分解:$\boldsymbol{A}=\boldsymbol{LDR}$,其中 $\boldsymbol{L}$ 为单位下三角阵,$\boldsymbol{D}$ 为对角阵,$\boldsymbol{R}$ 为单位上三角阵.

(2)$\boldsymbol{A}$ 有唯一的克洛脱(Crout)分解:$\boldsymbol{A}=\overline{\boldsymbol{L}}\,\overline{\boldsymbol{U}}$,其中 $\overline{\boldsymbol{L}}$ 为下三角阵,$\overline{\boldsymbol{U}}$ 为单位上三角阵.

求出一个矩阵的 $\boldsymbol{LU}$ 分解实际上就是要确定两个矩阵 $\boldsymbol{L}$ 和 $\boldsymbol{U}$ 中的元素,这时只要将一个可逆阵按照高斯消元法约化为一个上三角阵,则可得到矩阵的 $\boldsymbol{LU}$ 分解. 不过对于这种分解还可以通过下列直接的紧凑方式来完成.

设 $\boldsymbol{A}$ 为非奇异矩阵,且有分解式 $\boldsymbol{A}=\boldsymbol{LU}$,其中 $\boldsymbol{L}$ 为单位下三角阵,$\boldsymbol{U}$ 为上三角阵,即

$$\boldsymbol{A}=\begin{bmatrix}1 & & & \\ l_{21} & 1 & & \\ \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & & 1\end{bmatrix}\begin{bmatrix}u_{11} & u_{12} & \cdots & u_{1n} \\ & u_{22} & \cdots & u_{2n} \\ & & \ddots & \vdots \\ & & & u_{nn}\end{bmatrix}. \tag{6.16}$$

下面说明 $\boldsymbol{L}$ 和 $\boldsymbol{U}$ 的元素可以由 n 步直接计算定出,其中第 r 步定出 $\boldsymbol{U}$ 的第 r 行和 $\boldsymbol{L}$ 的第 r 列元素. 利用矩阵乘法,由式(6.16)有

$$a_{1i}=u_{1i}\quad (i=1,2,\cdots,n),\text{得 }\boldsymbol{U}\text{ 的第 1 行元素}.$$

$$a_{i1}=l_{i1}u_{11},l_{i1}=a_{i1}/u_{11}\quad (i=1,2,\cdots,n),\text{得 }\boldsymbol{L}\text{ 的第 1 列元素}.$$

设已经定出 $\boldsymbol{U}$ 的第 1 行到第 $r-1$ 行元素与 $\boldsymbol{L}$ 的第 1 列到第 $r-1$ 列元素. 根据式(6.16),利用矩阵乘法(当 $r<k$ 时,$l_{rk}=0$)有

$$a_{ri}=\sum_{k=1}^{n}l_{rk}u_{ki}=\sum_{k=1}^{r-1}l_{rk}u_{ki}+u_{ri},$$

故

$$u_{ri}=a_{ri}-\sum_{k=1}^{r-1}l_{rk}u_{ki}\quad (i=r,r+1,\cdots,n).$$

又由式(6.16)有

$$a_{ir}=\sum_{k=1}^{n}l_{ik}u_{kr}=\sum_{k=1}^{r-1}l_{ik}u_{kr}+l_{ir}u_{rr},$$

故

$$l_{ir}=\Big(a_{ir}-\sum_{k=1}^{r-1}l_{ik}u_{kr}\Big)/u_{rr}\quad (i=r+1,\cdots,n\text{ 且 }r\neq n).$$

一旦实现了矩阵 $\boldsymbol{A}$ 的 $\boldsymbol{LU}$ 分解,那么求解式(6.1)的问题就等价于求解如下两个三角形方程组:

(1)$\boldsymbol{Ly}=\boldsymbol{b}$,求 $\boldsymbol{y}$;

(2)$\boldsymbol{Ux}=\boldsymbol{y}$,求 $\boldsymbol{x}$.

这就相当于解方程组式(6.7)和式(6.8). 这时求解 $\boldsymbol{Ly}=\boldsymbol{b}$ 和 $\boldsymbol{Ux}=\boldsymbol{y}$ 的计算公式为

$$\begin{cases} y_1 = b_1, \\ y_i = b_i - \sum\limits_{k=1}^{i-1} l_{ik} y_k \quad (i = 2,3,\cdots,n); \end{cases} \tag{6.17}$$

$$\begin{cases} x_n = y_n / u_{nn}, \\ x_i = \left(y_i - \sum\limits_{k=i+1}^{n} u_{ik} x_k\right) / u_{ii} \quad (i = n-1, n-2, \cdots, 1). \end{cases} \tag{6.18}$$

当矩阵 $\boldsymbol{A}$ 为对称正定矩阵时,它的各阶顺序主子式不为零. 由定理 6.2 可知它存在唯一的 $\boldsymbol{LDR}$ 分解,即 $\boldsymbol{A}=\boldsymbol{LDR}$,其中 $\boldsymbol{L}$ 为单位下三角阵,$\boldsymbol{D}$ 为对角阵,$\boldsymbol{R}$ 为单位上三角阵.

由 $\boldsymbol{A}$ 的对称性可知 $\boldsymbol{LDR}=\boldsymbol{R}^{\mathrm{T}}\boldsymbol{DL}^{\mathrm{T}}$,由于分解式是唯一的,从而 $\boldsymbol{L}=\boldsymbol{R}^{\mathrm{T}}$,因此得到 $\boldsymbol{A}=\boldsymbol{LDL}^{\mathrm{T}}$. 设 $\boldsymbol{D}=\mathrm{diag}(d_1,d_2,\cdots,d_n)$,$d_i\neq 0$,$(i=1,2,\cdots,n)$. 下面进一步可以说明 $\boldsymbol{D}$ 的对角线元素均为正数,即 $d_i>0$.

由于 $\boldsymbol{L}$ 是单位下三角阵,所以 $\boldsymbol{L}^{\mathrm{T}}$ 是单位上三角阵,当然是非奇异矩阵. 故对于单位坐标向量 $\boldsymbol{e}_i$,存在非零向量 $\boldsymbol{x}_i$,使得

$$\boldsymbol{L}^{\mathrm{T}}\boldsymbol{x}_i = \boldsymbol{e}_i \quad (i = 1,2,\cdots,n).$$

另外

$$\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}_i = \boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{LDL}^{\mathrm{T}})\boldsymbol{x}_i = (\boldsymbol{L}^{\mathrm{T}}\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{D}(\boldsymbol{L}^{\mathrm{T}}\boldsymbol{x}_i) = \boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{D}\boldsymbol{e}_i = d_i$$

根据 $\boldsymbol{A}$ 是对称正定矩阵,则有 $\boldsymbol{x}_i^{\mathrm{T}}A\boldsymbol{x}_i>0$,从而 $d_i>0(i=1,2,\cdots,n)$. 这就说明了 $\boldsymbol{D}$ 的对角线元素全部是正数.

记

$$\boldsymbol{D}^{1/2} = \mathrm{diag}(\sqrt{d_1}, \sqrt{d_2}, \cdots, \sqrt{d_n}),$$

则有

$$\boldsymbol{A} = \boldsymbol{LDL}^{\mathrm{T}} = \boldsymbol{LD}^{1/2}\boldsymbol{D}^{1/2}\boldsymbol{L}^{\mathrm{T}} = (\boldsymbol{LD}^{1/2})(\boldsymbol{LD}^{1/2})^{\mathrm{T}} = \boldsymbol{L}_1\boldsymbol{L}_1^{\mathrm{T}}.$$

从而得到:

定理 6.3 如果 $\boldsymbol{A}$ 为 n 阶对称正定矩阵,则存在一个实的非奇异的下三角阵 $\boldsymbol{L}$ 使得 $\boldsymbol{A}=\boldsymbol{LL}^{\mathrm{T}}$. 当限定 $\boldsymbol{L}$ 的对角元素为正时,这种分解也是唯一的.

这种分解称为乔列斯基(Cholesky)分解.

若设

$$\boldsymbol{A} = \begin{bmatrix} l_{11} & & & \\ l_{21} & l_{22} & & \\ \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & \cdots & l_{n1} \\ & l_{22} & \cdots & l_{n2} \\ & & \ddots & \\ & & & l_{nn} \end{bmatrix}$$

其中 $l_{ii}>0(i=1,2,\cdots,n)$. 由矩阵乘法及 $l_{jk}=0$(当 $j<k$ 时),利用矩阵乘法容易得到

$$a_{ij}=\sum_{k=1}^{n}l_{ik}l_{jk}=\sum_{k=1}^{j-1}l_{ik}l_{jk}+l_{jj}l_{ij}.$$

从而得到解对称正定方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 的平方根法计算公式:

对于 $j=1,2,\cdots,n$

步骤 1:$l_{jj}=\left(a_{jj}-\sum_{k=1}^{j-1}l_{jk}^2\right)^{\frac{1}{2}}$.

步骤 2:$l_{ij}=\left(a_{ij}-\sum_{k=1}^{j-1}l_{ik}l_{jk}\right)/l_{jj}\quad(i=j+1,\cdots,n)$,

求解 $\boldsymbol{Ax}=\boldsymbol{b}$,即求解两个三角形方程组:(1)$\boldsymbol{Ly}=\boldsymbol{b}$,求 $\boldsymbol{y}$;(2)$\boldsymbol{L}^{\mathrm{T}}\boldsymbol{x}=\boldsymbol{y}$,求 $\boldsymbol{x}$.

步骤 3:$y_i=\left(b_i-\sum_{k=1}^{i-1}l_{ik}y_k\right)/l_{ii}\quad(i=1,2,\cdots,n)$.

步骤 4:$x_i=\left(b_i-\sum_{k=i+1}^{n}l_{ki}x_k\right)/l_{ii}\quad(i=n,n-1,\cdots,1)$.

在引言的例 6.3 中,出现了一个对角线占优的三对角方程组,实际上在微分方程数值解中经常出现这种类型的方程组. 利用矩阵分解可以得出非常简便的求解方法称为追赶法.

设有方程组

$$\begin{bmatrix}b_1 & c_1 & & & & \\ a_2 & b_2 & c_2 & & & \\ \ddots & \ddots & \ddots & & & \\ & a_i & b_i & c_i & & \\ & \ddots & \ddots & \ddots & & \\ & & a_{n-1} & b_{n-1} & c_{n-1} & \\ & & & a_n & b_n & \end{bmatrix}\begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix}=\begin{bmatrix}f_1\\f_2\\\vdots\\f_n\end{bmatrix},\tag{6.19}$$

简记为 $\boldsymbol{Ax}=\boldsymbol{f}$. 其中当 $|i-j|>1$ 时,$a_{ij}=0$,并且

(1)$|b_1|>|c_1|>0$;

(2)$|b_i|\geqslant|a_i|+|c_i|$,$a_i$、$c_i\neq0\quad(i=2,3,\cdots,n-1)$;

(3)$|b_n|>|a_n|>0$.

利用矩阵的直接三角分解法来推导解三对角线方程组式(6.19)的计算公式.

由系数阵 $\boldsymbol{A}$ 的特点,可以将 $\boldsymbol{A}$ 分解为两个三角阵的乘积

$$\boldsymbol{A}=\boldsymbol{LU},$$

其中 $\boldsymbol{L}$ 为下三角矩阵,$\boldsymbol{U}$ 为单位上三角矩阵. 下面来说明这种分解是可能的. 设

$$\boldsymbol{A}=\begin{bmatrix} b_1 & c_1 & & & \\ a_2 & b_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & c_{n-1} \\ & & & a_n & b_n \end{bmatrix}=\begin{bmatrix} \alpha_1 & & & \\ r_2 & \alpha_2 & & \\ & \ddots & \ddots & \\ & & \ddots & \alpha_n \\ & & & r_n \end{bmatrix}\begin{bmatrix} 1 & \beta_1 & & & \\ & 1 & \beta_2 & & \\ & & \ddots & \ddots & \\ & & & \ddots & \beta_{n-1} \\ & & & & 1 \end{bmatrix},$$

(6.20)

其中 α_i、β_i、r_i 为待定系数．利用矩阵乘法不难得到：

(1)$\beta_1=c_1/b_1$，$\beta_i=c_i/(b_i-a_i\beta_{i-1})$　$(i=2,3,\cdots,n-1)$；

(2)$\alpha_1=b_1$，$r_i=a_i$，$\alpha_i=b_i-a_i\beta_{i-1}$　$(i=2,3,\cdots,n-1)$.

求解 $\boldsymbol{Ax}=\boldsymbol{f}$ 等价于求解以下两个三角形方程组：

(1)$\boldsymbol{Ly}=\boldsymbol{f}$，求 $\boldsymbol{y}$；

(2)$\boldsymbol{Ux}=\boldsymbol{y}$，求 $\boldsymbol{x}$.

解　(1)$\boldsymbol{Ly}=\boldsymbol{f}$，

$y_1=f_1/b_1$，

$y_i=(f_i-a_iy_{i-1})/(b_i-a_i\beta_{i-1})$　$(i=2,3,\cdots,n)$；

(2)$\boldsymbol{Ux}=\boldsymbol{y}$，

$x_n=y_n$，

$x_i=y_i-\beta_ix_{i+1}$　$(i=n-1,n-2,\cdots,2,1)$.

将计算系数 $\beta_1\to\beta_2\to\cdots\to\beta_{n-1}$ 及 $y_1\to y_2\to\cdots\to y_n$ 的过程称为追的过程，将计算方程组的解 $x_n\to x_{n-1}\to\cdots\to x_1$ 的过程称为赶的过程．

6.3　向量和矩阵的范数

为了对线性空间中的元素大小进行衡量，需要引进范数的定义，它是空间 $\boldsymbol{R}^n$ 中向量长度概念的直接推广．范数是泛函分析中的重要概念，在数学学科及其他学科中的应用非常广泛，下面给出它们的公理化定义．

定义 6.1　设 $\boldsymbol{X}$ 为线性空间，对于任意 $\boldsymbol{x}\in\boldsymbol{X}$，存在唯一的一个实数 $\|\boldsymbol{x}\|$ 与之对应，且满足：

(1) $\|\boldsymbol{x}\|\geqslant 0$，$\|\boldsymbol{x}\|=0$ 当且仅当 $x=0$；

(2) $\|\alpha\boldsymbol{x}\|=|\alpha|\|\boldsymbol{x}\|$，$\forall\alpha\in\boldsymbol{R}$；

(3) $\|\boldsymbol{x}+\boldsymbol{y}\|\leqslant\|\boldsymbol{x}\|+\|\boldsymbol{y}\|$，$\forall\boldsymbol{x},\boldsymbol{y}\in\boldsymbol{X}$.

则称$(X,\|\cdot\|)$为赋范线性空间，简称为赋范空间，也称 X 为赋范空间．$\|\boldsymbol{x}-\boldsymbol{y}\|$ 表示元素 $\boldsymbol{x}$、$\boldsymbol{y}$ 之间的距离，经常用来衡量两个元素之间的误差大小．

例 6.6　$\boldsymbol{X}=\boldsymbol{R}^n$，$\forall\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$，定义三种常用的范数：

(1) $\|\boldsymbol{x}\|_\infty=\max\limits_{1\leqslant i\leqslant n}|x_i|$，称为最大范数或$\infty$－范数；

(2) $\|\boldsymbol{x}\|_1=\sum\limits_{i=1}^{n}|x_i|$ 称为 1－范数；

(3) $\|\boldsymbol{x}\|_2=\left(\sum\limits_{i=1}^{n}x_i^2\right)^{\frac{1}{2}}$，称为欧氏范数或 2－范数．

可以验证上述三种范数定义均满足范数公理．

例 6.7 $\boldsymbol{X}=\boldsymbol{C}[a,b]$，$C[a,b]$是闭区间$[a,b]$上所有的连续函数组成的连续函数空间．$\forall f\in C[a,b]$，定义三种常用的范数：

(1) $\|f\|_\infty=\max\limits_{a\leqslant x\leqslant b}|f(x)|$，称为最大范数或∞－范数；

(2) $\|f\|_1=\int_a^b|f(x)|\mathrm{d}x$，称为 1－范数；

(3) $\|f\|_2=\left(\int_a^b f^2(x)\mathrm{d}x\right)^{\frac{1}{2}}$，称为 2－范数．

可以验证上述三种范数均满足范数公理．

例 6.8 $\boldsymbol{X}=\boldsymbol{R}^{n\times n}$，$\boldsymbol{R}^{n\times n}$表示所有的 n 阶方阵组成的线性空间，对于任意 n 阶方阵 $\boldsymbol{A}\in\boldsymbol{R}^{n\times n}$，也可以定义范数．不过对于矩阵的范数，除了要求满足定义 6.1 中的三条之外，一般还要要求

$$\|\boldsymbol{AB}\|\leqslant\|\boldsymbol{A}\|\ \|\boldsymbol{B}\|.$$

对一个方阵 $\boldsymbol{A}$，定义三种常用的矩阵范数：

(1) $\|\boldsymbol{A}\|_\infty=\max\limits_{1\leqslant i\leqslant n}\sum\limits_{i=1}^{n}|a_{ij}|$，称为 $\boldsymbol{A}$ 的行范数；

(2) $\|\boldsymbol{A}\|_1=\max\limits_{1\leqslant j\leqslant n}\sum\limits_{i=1}^{n}|a_{ij}|$，称为 $\boldsymbol{A}$ 的列范数；

(3) $\|\boldsymbol{A}\|_2=\sqrt{\lambda_{\max}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})}$，称为 $\boldsymbol{A}$ 的 2－范数．

其中 $\lambda_{\max}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})$表示矩阵 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ 的最大特征值．

可以验证上述三种范数均满足范数公理及 $\|\boldsymbol{AB}\|\leqslant\|\boldsymbol{A}\|\ \|\boldsymbol{B}\|$．还可以定义一种范数称之为 Frobenius 范数，即

$$\|\boldsymbol{A}\|_{\mathrm{F}}=\sqrt{\sum_{i,j=1}^{n}a_{ij}^2}.$$

由于矩阵可以和向量做乘法，而矩阵和向量都可以定义范数，在误差估计中经常希望矩阵的某种范数和向量的某种范数能满足一个所谓的相容性质：

$$\|\boldsymbol{Ax}\|\leqslant\|\boldsymbol{A}\|\cdot\|\boldsymbol{x}\|.$$

就上述常用范数而言，相容性质均成立，即

$$\|\boldsymbol{Ax}\|_1\leqslant\|\boldsymbol{A}\|_1\|\boldsymbol{x}\|_1,\ \|\boldsymbol{Ax}\|_\infty\leqslant\|\boldsymbol{A}\|_\infty\|\boldsymbol{x}\|_\infty,\ \|\boldsymbol{Ax}\|_2\leqslant\|\boldsymbol{A}\|_2\|\boldsymbol{x}\|_2.$$

例 6.9 设 $\boldsymbol{A}=\begin{pmatrix}1&-2\\-3&4\end{pmatrix}$，计算 $\boldsymbol{A}$ 的各种范数．

解 $\|\boldsymbol{A}\|_1=6$，$\|\boldsymbol{A}\|_\infty=7$，$\|\boldsymbol{A}\|_2=\sqrt{15+\sqrt{221}}\approx5.46$.

定义 6.2 设$\{\boldsymbol{x}^{(k)}\}$为 $\boldsymbol{R}^n$ 中一向量序列，$\boldsymbol{x}^*\in\boldsymbol{R}^n$，记 $\boldsymbol{x}^{(k)}=(x_1^{(k)},x_2^{(k)},\cdots,x_n^{(k)})^{\mathrm{T}}$，$\boldsymbol{x}^*=(x_1^*,\cdots,x_n^*)^{\mathrm{T}}$．如果 $\lim\limits_{k\to\infty}x_i^{(k)}=x_i^*\ (x=1,2,3,\cdots,n)$，则称 $\boldsymbol{x}^{(k)}$ 收敛于向量 $\boldsymbol{x}^*$ 记为

$$\lim_{k\to\infty}\boldsymbol{x}^{(k)}=\boldsymbol{x}^*.$$

定理 6.4 $\lim\limits_{k\to\infty}\boldsymbol{x}^{(k)}=\boldsymbol{x}^{*}\Leftrightarrow\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{*}\|\to 0$，其中 $\|\cdot\|$ 为向量的任一种范数.

定义 6.3 设 $\boldsymbol{A}\in\boldsymbol{R}^{n\times n}$ 的特征值为 $\lambda_i(i=1,2,\cdots,n)$，称 $\rho(\boldsymbol{A})=\max\limits_{1\leqslant i\leqslant n}|\lambda_i|$ 为 $\boldsymbol{A}$ 的谱半径.

定理 6.5 设 $\boldsymbol{A}\in\boldsymbol{R}^{n\times n}$，则 $\rho(\boldsymbol{A})\leqslant\|\boldsymbol{A}\|$，即 $\boldsymbol{A}$ 的谱半径不超过 $\boldsymbol{A}$ 的任何一种算子范数.

6.4 矩阵的条件数

在引言中的例 6.2 的求解是比较奇怪的，右端常数项只发生了微小的变化，而解却发生了重大的变化. 产生这种情况的原因和系数矩阵有关，下面给出一个病态矩阵的定义，并讨论如何刻画矩阵或方程组的病态性质.

定义 6.4 如果矩阵 $\boldsymbol{A}$ 或常数项 $\boldsymbol{b}$ 的微小变化，引起方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 解的巨大变化，则称此方程组为“病态”方程组，矩阵 $\boldsymbol{A}$ 称为“病态”矩阵（相对于方程组而言），否则称方程组为“良态”方程组，$\boldsymbol{A}$ 称为“良态”矩阵.

应该注意，矩阵的“病态”性质是矩阵本身的特性，下面找出刻画矩阵“病态”性质的量. 设有方程组

$$\boldsymbol{Ax}=\boldsymbol{b}, \tag{6.21}$$

其中 $\boldsymbol{A}$ 为非奇异阵，$\boldsymbol{x}$ 为式(6.21)的精确解. 以下分析当方程组的系数矩阵 $\boldsymbol{A}$（或 $\boldsymbol{b}$）有微小误差（扰动）时对解的影响.

现设 $\boldsymbol{A}$ 是精确的，$\boldsymbol{b}$ 有误差 $\delta\boldsymbol{b}$，解为 $\boldsymbol{x}+\delta\boldsymbol{x}$，则

$$\boldsymbol{A}(\boldsymbol{x}+\delta\boldsymbol{x})=\boldsymbol{b}+\delta\boldsymbol{b},\delta x=\boldsymbol{A}^{-1}\delta\boldsymbol{b},$$

$$\|\delta\boldsymbol{x}\|\leqslant\|\boldsymbol{A}^{-1}\|\,\|\delta\boldsymbol{b}\|. \tag{6.22}$$

由式(6.21)有

$$\|\boldsymbol{b}\|\leqslant\|\boldsymbol{A}\|\,\|\boldsymbol{x}\|,$$

$$\frac{1}{\|\boldsymbol{x}\|}\leqslant\frac{\|\boldsymbol{A}\|}{\|\boldsymbol{b}\|}\quad(\text{设}\ \boldsymbol{b}\neq 0) \tag{6.23}$$

从而得到

$$\frac{\|\delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|}\leqslant\|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{A}\|\,\frac{\|\delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}.$$

上式给出了解的相对误差的上界，常数项 $\boldsymbol{b}$ 的相对误差在解中可能放大 $\|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{A}\|$ 倍.

现设 $\boldsymbol{b}$ 是精确的，$\boldsymbol{A}$ 有微小误差（扰动）$\delta\boldsymbol{A}$，解为 $\boldsymbol{x}+\delta\boldsymbol{x}$，类似地可以说明量 $\|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{A}\|$ 越小，由 $\boldsymbol{A}$ 的相对误差引起的解的相对误差就越小；量 $\|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{A}\|$ 越大，解的相对误差就可能越大.

因此量 $\|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{A}\|$ 实际上刻画了解对原始数据变化的灵敏程度，即刻画了方程组的“病态”程度，于是引进下述定义.

定义 6.5 设 $\boldsymbol{A}$ 为非奇异矩阵，称数 $\mathrm{cond}(\boldsymbol{A})_{\upsilon}=\|\boldsymbol{A}^{-1}\|_{\upsilon}\|\boldsymbol{A}\|_{\upsilon}(\upsilon=1,2\ \text{或}\infty)$ 为矩阵的条件数. 习惯上当 $\boldsymbol{A}$ 为不可逆阵时，定义 $\mathrm{cond}(\boldsymbol{A})=\infty$.

矩阵的条件数是一个十分重要的概念，它和矩阵的范数有关系. 由上面讨论可知，当 $\boldsymbol{A}$

的条件数相对的大即 $\mathrm{cond}(\boldsymbol{A})\gg 1$ 时，则式(6.21)是“病态”的(即 $\boldsymbol{A}$ 是“病态”矩阵，或者说 $\boldsymbol{A}$ 是坏条件的)；当 $\boldsymbol{A}$ 的条件数相对的小，则式(6.21)是“良态”的(或者说 $\boldsymbol{A}$ 是好条件的)．$\boldsymbol{A}$ 的条件数越大，方程组的病态程度越严重，也就越难得到方程组的比较准确的解．

通常使用的条件数，有

(1)$\mathrm{cond}(\boldsymbol{A})_\infty=\|\boldsymbol{A}^{-1}\|_\infty\|\boldsymbol{A}\|_\infty$.

(2)$\boldsymbol{A}$ 的谱条件数

$$\mathrm{cond}(\boldsymbol{A})_2=\|\boldsymbol{A}\|_2\|\boldsymbol{A}^{-1}\|_2=\sqrt{\frac{\lambda_{\max}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})}{\lambda_{\min}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})}}\ .$$

当 $\boldsymbol{A}$ 为对称矩阵时，$\mathrm{cond}(\boldsymbol{A})_2=\dfrac{|\lambda_1|}{|\lambda_n|}$，其中 λ_1,λ_n 为 $\boldsymbol{A}$ 的绝对值最大和绝对值最小的特征值．

条件数的性质如下：

(1)对任何非奇异矩阵 $\boldsymbol{A}$ 都有 $\mathrm{cond}(\boldsymbol{A})_v\geqslant 1$. 事实上，

$$\mathrm{cond}(\boldsymbol{A})_v=\|\boldsymbol{A}^{-1}\|_v\|\boldsymbol{A}\|_v\geqslant\|\boldsymbol{A}^{-1}\boldsymbol{A}\|_v=1\ .$$

(2)设 $\boldsymbol{A}$ 为非奇异矩阵且 $c\neq 0$(常数)，则 $\mathrm{cond}(c\boldsymbol{A})_v=\mathrm{cond}(\boldsymbol{A})_v$．

(3)如果 $\boldsymbol{A}$ 为正交矩阵，则 $\mathrm{cond}(\boldsymbol{A})_2=1$；如果 $\boldsymbol{A}$ 为非奇异矩阵，$\boldsymbol{R}$ 为正交矩阵，则 $\mathrm{cond}(\boldsymbol{RA})_2=\mathrm{cond}(\boldsymbol{AR})_2=\mathrm{cond}(\boldsymbol{A})_2$.

例 6.10 已知 Hilbert 矩阵

$$\boldsymbol{H}_n=\begin{bmatrix}1&\dfrac{1}{2}&\cdots&\dfrac{1}{n}\\[2ex]\dfrac{1}{2}&\dfrac{1}{3}&\cdots&\dfrac{1}{n+1}\\[2ex]\vdots&\vdots&&\vdots\\[2ex]\dfrac{1}{n}&\dfrac{1}{n+1}&\cdots&\dfrac{1}{2n-1}\end{bmatrix}$$

计算 $\boldsymbol{H}_3$ 的条件数．

解

$$\boldsymbol{H}_3=\begin{bmatrix}1&\dfrac{1}{2}&\dfrac{1}{3}\\[2ex]\dfrac{1}{2}&\dfrac{1}{3}&\dfrac{1}{4}\\[2ex]\dfrac{1}{3}&\dfrac{1}{4}&\dfrac{1}{5}\end{bmatrix},\quad \boldsymbol{H}_3^{-1}=\begin{bmatrix}9&-36&30\\-36&192&-180\\30&-180&180\end{bmatrix}.$$

(1)计算 $\boldsymbol{H}_3$ 条件数 $\mathrm{cond}(H_3)_\infty$.

$\|\boldsymbol{H}_3\|_\infty=11/6$，$\|\boldsymbol{H}_3^{-1}\|_\infty=408$，所以 $\mathrm{cond}(\boldsymbol{H}_3)_\infty=748$. 同样可计算出 $\mathrm{cond}(\boldsymbol{H}_6)_\infty=2.9\times 10^7$，一般 $\boldsymbol{H}_n$ 矩阵当 n 越大时，病态越严重．

(2)考虑方程组 $\boldsymbol{H}_3x=(11/6,13/12,47/60)^{\mathrm{T}}=\boldsymbol{b}$.

设 $\boldsymbol{H}_3$ 及 $\boldsymbol{b}$ 有微小误差(取 3 位有效数字)有

$$\begin{bmatrix} 1.000 & 0.500 & 0.333 \\ 0.500 & 0.333 & 0.250 \\ 0.333 & 0.250 & 0.200 \end{bmatrix} \begin{bmatrix} x_1+\delta x_1 \\ x_2+\delta x_2 \\ x_3+\delta x_3 \end{bmatrix} = \begin{bmatrix} 1.83 \\ 1.08 \\ 0.783 \end{bmatrix}, \tag{6.24}$$

简记为$(\boldsymbol{H}_3+\delta\boldsymbol{H}_3)(\boldsymbol{x}+\delta\boldsymbol{x})=\boldsymbol{b}+\delta\boldsymbol{b}$. 方程组$\boldsymbol{H}_3\boldsymbol{x}=\boldsymbol{b}$与式(6.24)的精确解分别为

$$\boldsymbol{x}=(1,1,1)^{\mathrm{T}},\boldsymbol{x}+\delta\boldsymbol{x}=(1.089512538,0.487967062,1.491002798)^{\mathrm{T}}.$$

于是

$$\delta\boldsymbol{x}=(0.0895,-0.5120,0.4910)^{\mathrm{T}},$$

$$\frac{\|\delta\boldsymbol{H}_3\|_\infty}{\|\boldsymbol{H}_3\|_\infty}\approx 0.18\times 10^{-3}<0.02\%,$$

$$\frac{\|\delta\boldsymbol{b}\|_\infty}{\|\boldsymbol{b}\|_\infty}\approx 0.182\%,$$

$$\frac{\|\delta\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty}\approx 51.2\%.$$

这就是说$\boldsymbol{H}_3$与$\boldsymbol{b}$相对误差不超过0.2%，而引起解的相对误差却超过50%.

在曲线拟合中，若拟合曲线的基函数选取为$1,x,x^2,\cdots,x^{n-1}$，权函数取为1，则得到的法方程的系数矩阵就是Hilbert矩阵. 由此看来，Hilbert矩阵高度病态. 在曲线拟合中，为了避免求解病态方程组，可以选取正交多项式作为基函数.

对于病态方程组的求解，一般可采用高精度的算术运算或者采用预处理的方法.

习　题

1. 用高斯消去法求解下列线性代数方程组

$$\begin{cases} 6x_1-2x_2+2x_3+4x_4=16, \\ 12x_1-8x_2+6x_3+10x_4=26, \\ 3x_1-13x_2+9x_3+3x_4=-19, \\ -6x_1+4x_2+x_3-18x_4=-34. \end{cases}$$

2. 用高斯列主元消去法求解下列线性代数方程组

$$\begin{cases} x_1+3x_2-2x_3-4x_4=3, \\ 2x_1+6x_2-7x_3-10x_4=-2, \\ -x_1-x_2+5x_3+9x_4=-14, \\ -3x_1-5x_2+15x_4=-6. \end{cases}$$

3. 用追赶法求解三对角方程组$\boldsymbol{A}x=\boldsymbol{b}$，其中

$$\boldsymbol{A}=\begin{pmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{pmatrix}, \boldsymbol{b}=\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

4. 对于 2 中的线性代数方程组，分别求系数矩阵的 1－范数、2－范数和∞－范数；对于该系数矩阵，分别求矩阵第一个列向量的 1－范数、2－范数和∞－范数.

5. 设 $\boldsymbol{A}$ 为 n 阶实对称阵，定义

$$\| x \|_{\boldsymbol{A}} = \sqrt{(\boldsymbol{A}x, x)},$$

证明 $\| x \|_{\boldsymbol{A}}$ 是 $\boldsymbol{R}^n$ 上向量的一种范数.

6. 对于向量 $\boldsymbol{x} \in \boldsymbol{R}^n$，证明

$$\| \boldsymbol{x} \|_\infty \leqslant \| \boldsymbol{x} \|_1 \leqslant n \| \boldsymbol{x} \|_\infty .$$

7 解线性方程组的迭代法

7.1 引言

考虑线性方程组

$$\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b} \tag{7.1}$$

其中 $\boldsymbol{A}$ 为非奇异矩阵，当 $\boldsymbol{A}$ 为低阶稠密矩阵时，第 6 章所讨论的高斯列主元消去法是解式(7.1)的有效方法，或者通过对系数矩阵的分解从而求解两个三角形方程组．但是，对于由工程技术中产生的大型稀疏矩阵方程组($\boldsymbol{A}$ 的阶数 n 很大，但零元素较多，例如，$n\geqslant 10^4$，由某些偏微分方程数值解所产生的线性方程组)，利用迭代法求解式(7.1)是合适的．迭代法用在计算机内存和运算两方面，通常都可利用 $\boldsymbol{A}$ 中有大量零元素的特点．

本章将介绍迭代法的一些基本理论及雅可比迭代法、高斯—塞德尔迭代法、超松弛迭代法，而超松弛迭代法应用很广泛．

下面首先举一个简单的例子，以便了解迭代法的思想．

例 7.1 求解方程组

$$\begin{cases}8x_1-3x_2+2x_3=20,\\ 4x_1+11x_2-x_3=33,\\ 6x_1+3x_2+12x_3=36.\end{cases} \tag{7.2}$$

解 记为 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$，其中

$$\boldsymbol{A}=\begin{bmatrix}8 & -3 & 2\\ 4 & 11 & -1\\ 6 & 3 & 12\end{bmatrix},\boldsymbol{x}=\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix},\boldsymbol{b}=\begin{bmatrix}20\\ 33\\ 36\end{bmatrix}.$$

方程组的精确解是 $\boldsymbol{x}^*=(3,2,1)^{\mathrm{T}}$．现将式(7.2)改写成下面等价的方程组

$$\begin{cases}x_1=\dfrac{1}{8}(3x_2-2x_3+20),\\ x_2=\dfrac{1}{11}(-4x_1+x_3+33),\\ x_3=\dfrac{1}{12}(-6x_1-3x_2+36),\end{cases} \tag{7.3}$$

或写为 $\boldsymbol{x}=\boldsymbol{B}_0\boldsymbol{x}+\boldsymbol{f}$，其中

$$\boldsymbol{B}_0=\begin{bmatrix}0 & \frac{3}{8} & -\frac{2}{8}\\ -\frac{4}{11} & 0 & \frac{1}{11}\\ -\frac{6}{12} & -\frac{3}{12} & 0\end{bmatrix},\boldsymbol{f}=\begin{bmatrix}\frac{20}{8}\\ \frac{33}{11}\\ \frac{36}{12}\end{bmatrix}.$$

任取初始值，例如取 $\boldsymbol{x}^{(0)}=(0,0,0)^{\mathrm{T}}$. 将这些值代入式(7.3)右边(若式(7.3)为等式即求得方程组的解，但一般不满足)，得到新的值

$$\boldsymbol{x}^{(1)}=(x_1^{(1)},x_2^{(1)},x_3^{(1)})^{\mathrm{T}}=(2.5,3,3)^{\mathrm{T}}.$$

再将 $\boldsymbol{x}^{(1)}$ 分量代入式(7.3)右边得到 $\boldsymbol{x}^{(2)}$，反复利用这个计算程序，得到一向量序列和一般的计算公式(迭代公式)

$$\boldsymbol{x}^{(0)}=\begin{bmatrix}x_1^{(0)}\\ x_2^{(0)}\\ x_3^{(0)},\end{bmatrix},\boldsymbol{x}^{(1)}=\begin{bmatrix}x_1^{(1)}\\ x_2^{(1)}\\ x_3^{(1)},\end{bmatrix},\cdots,\boldsymbol{x}^{(k)}=\begin{bmatrix}x_1^{(k)}\\ x_2^{(k)}\\ x_3^{(k)}\end{bmatrix},\cdots$$

$$\begin{cases}x_1^{(k+1)}=(3x_2^{(k)}-2x_3^{(k)}+20)/8,\\ x_2^{(k+1)}=(-4x_1^{(k)}+x_3^{(k)}+33)/11,\\ x_3^{(k+1)}=(-6x_1^{(k)}-3x_2^{(k)}+36)/12.\end{cases}\tag{7.4}$$

简写为

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{B}_0\boldsymbol{x}^{(k)}+\boldsymbol{f},$$

其中，k 表示迭代次数($k=0,1,2,\cdots$).

迭代到第 10 次有

$$\boldsymbol{x}^{(10)}=(3.000032,1.999838,0.9998813)^{\mathrm{T}},$$

$$\|\boldsymbol{\varepsilon}^{(10)}\|_\infty=0.000187.$$

其中，$\boldsymbol{\varepsilon}^{(10)}=\boldsymbol{x}^{(10)}-\boldsymbol{x}^*$ 表示迭代值和准确值之间的误差向量.

从此例可以看出，由迭代法做出的向量序列 $\boldsymbol{x}^{(k)}$ 逐步逼近方程组的精确解 $\boldsymbol{x}^*$.

对于任何一个方程组 $\boldsymbol{Ax}=\boldsymbol{b}$，如何运用迭代法构造迭代序列以逼近方程组的解呢?

第一步：将方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 按照某种方式转化成等价形式的方程组

$$\boldsymbol{x}=\boldsymbol{Bx}+\boldsymbol{f}.$$

第二步：任取初始迭代向量 $\boldsymbol{x}^{(0)}$，按照下述迭代公式构造向量序列

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{Bx}^{(k)}+\boldsymbol{f}\quad(k=0,1,2,\cdots)\tag{7.5}$$

其中，$\boldsymbol{B}$ 称为迭代矩阵，k 表示迭代次数.

第三步：讨论向量序列 $\boldsymbol{x}^{(k+1)}$ 是否收敛.

定义 7.1 (1)对于给定的方程组 $\boldsymbol{x}=\boldsymbol{B}\boldsymbol{x}+\boldsymbol{f}$，用式(7.5)逐步代入求近似解的方法称为迭代法(或称为一阶定常迭代法，这里 $\boldsymbol{B}$ 与 k 无关).

(2)如果 $\lim\limits_{k\to\infty}x^{(k)}$ 存在(记为 $\boldsymbol{x}^*$)，称此迭代法收敛，显然 $\boldsymbol{x}^*$ 就是方程组的解，否则称此迭代法发散.

由上述讨论，需要研究向量序列 $\{\boldsymbol{x}^{(k)}\}$ 的收敛性. 设 $\boldsymbol{x}^*$ 是方程组 $\boldsymbol{x}=\boldsymbol{B}\boldsymbol{x}+\boldsymbol{f}$ 唯一的准确解，即

$$\boldsymbol{x}^* = \boldsymbol{B}\boldsymbol{x}^* + \boldsymbol{f}. \tag{7.6}$$

引进误差向量

$$\boldsymbol{\varepsilon}^{(k+1)} = \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^*, \tag{7.7}$$

由式(7.5)减去式(7.6)，得

$$\boldsymbol{\varepsilon}^{(k+1)} = \boldsymbol{B}\boldsymbol{\varepsilon}^{(k)} \quad (k=0,1,2,\cdots),$$

递推得

$$\boldsymbol{\varepsilon}^{(k)} = \boldsymbol{B}\boldsymbol{\varepsilon}^{(k-1)} = \cdots = \boldsymbol{B}^k\boldsymbol{\varepsilon}^{(0)}. \tag{7.8}$$

由式(7.8)可知，要考察 $\{\boldsymbol{x}^{(k)}\}$ 的收敛性，就要研究 $\boldsymbol{B}$ 在什么条件下有 $\boldsymbol{\varepsilon}^{(k)}\to 0\ (k\to\infty)$，亦即要研究 $\boldsymbol{B}$ 满足什么条件时有 $\boldsymbol{B}^k\to\boldsymbol{O}$(零矩阵)$(k\to\infty)$.

7.2 基本迭代法

考虑 n 阶线性代数方程组 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$，即

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2, \\ \qquad\qquad \cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n. \end{cases} \tag{7.9}$$

这里系数矩阵 $\boldsymbol{A}=(a_{ij})_{n\times n}$ 是非奇异矩阵，$\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$，$\boldsymbol{b}=(b_1,b_2,\cdots,b_n)^{\mathrm{T}}$，并且 $a_{ii}\neq 0$. 为了讨论方便，首先将系数矩阵 $\boldsymbol{A}$ 分裂成

$$\boldsymbol{A} = \boldsymbol{D} - \boldsymbol{L} - \boldsymbol{U} \tag{7.10}$$

其中

$$\boldsymbol{D} = \begin{bmatrix} a_{11} & & & \\ & a_{22} & & \\ & & \ddots & \\ & & & a_{nn} \end{bmatrix}$$

$$\boldsymbol{L}=-\begin{bmatrix}0 & & & \\ a_{21} & 0 & & \\ a_{31} & a_{32} & 0 & \\ \vdots & \vdots & & \ddots \\ a_{n1} & a_{n2} & \cdots & 0\end{bmatrix},\ \boldsymbol{U}=-\begin{bmatrix}0 & a_{12} & a_{13} & \cdots & a_{1n} \\ & 0 & a_{23} & \cdots & a_{2n} \\ & & \ddots & & \vdots \\ & & & & 0\end{bmatrix}$$

7.2.1 雅可比迭代法

从式(7.9)第 i 个方程($i=1,2,\cdots,n$)解出 x_i,得到与式(7.9)等价的方程组

$$\begin{cases}x_1=\dfrac{1}{a_{11}}(b_1-a_{12}x_2-a_{13}x_3-\cdots-a_{1n}x_n),\\ x_2=\dfrac{1}{a_{22}}(b_2-a_{21}x_1-a_{23}x_3-\cdots-a_{2n}x_n),\\ \cdots\cdots\\ x_n=\dfrac{1}{a_{nn}}(b_n-a_{n1}x_1-a_{n2}x_2-\cdots-a_{n,n-1}x_{n-1}),\end{cases}\tag{7.11}$$

由此,建立迭代格式

$$\begin{cases}x_1^{(k+1)}=\dfrac{1}{a_{11}}(b_1-a_{12}x_2^{(k)}-a_{13}x_3^{(k)}-\cdots-a_{1n}x_n^{(k)}),\\ x_2^{(k+1)}=\dfrac{1}{a_{22}}(b_2-a_{21}x_1^{(k)}-a_{23}x_3^{(k)}-\cdots-a_{2n}x_n^{(k)}),\\ \cdots\cdots\\ x_n^{(k+1)}=\dfrac{1}{a_{nn}}(b_n-a_{n1}x_1^{(k)}-a_{n2}x_2^{(k)}-\cdots-a_{n,n-1}x_{n-1}^{(k)}).\end{cases}\tag{7.12}$$

给出初值 $\boldsymbol{x}^{(0)}=(x_1^{(0)},x_2^{(0)},\cdots,x_n^{(0)})^{\mathrm{T}}$ 后,首先令 $k=0$,由式(7.12)的第一式计算出 $x_1^{(1)}$,由式(7.12)的第二式计算出 $x_2^{(1)}$,依次类推计算出 $x_n^{(1)}$,这就完成了第一次迭代过程. 然后令 $k=1$,依次计算出 $x_1^{(2)},x_2^{(2)},\cdots,x_n^{(2)}$ 完成第二次迭代过程,依次类推完成第 k 次迭代过程,得到一个向量序列$\{\boldsymbol{x}^{(k)}\}$.

若 $\boldsymbol{x}^{(k)}\to\boldsymbol{x}^*$,则 $\boldsymbol{x}^*$ 是原来线性代数方程组式(7.11)的准确解. 这种迭代方法称为雅可比(Jacobi)迭代法. 式(7.12)是雅可比迭代法的分量形式.

如果借助于求和符号,则式(7.12)可以写成更加简洁的形式

$$x_i^{(k+1)}=\frac{1}{a_{ii}}\Big(b_i-\sum_{j=1}^{i-1}a_{ij}x_j^{(k)}-\sum_{j=i+1}^{n}a_{ij}x_j^{(k)}\Big)\quad(i=1,2,\cdots,n).\tag{7.13}$$

为了判断雅可比迭代是否收敛,需要将方程组转化为 $\boldsymbol{x}=\boldsymbol{B}\boldsymbol{x}+\boldsymbol{f}$ 的形式,也就是说需要求出雅可比迭代矩阵 $\boldsymbol{B}$. 利用式(7.10)及矩阵乘法,式(7.12)可以改写成如下的矩阵形式

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{U})\boldsymbol{x}^{(k)} + \boldsymbol{D}^{-1}\boldsymbol{b}. \tag{7.14}$$

记

$$\boldsymbol{B}_{\mathrm{J}} = \boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{U}) = \boldsymbol{D}^{-1}(\boldsymbol{D}-\boldsymbol{A}) = \boldsymbol{E}-\boldsymbol{D}^{-1}\boldsymbol{A}, \boldsymbol{f}_{\mathrm{J}} = \boldsymbol{D}^{-1}\boldsymbol{b},$$

其中，$\boldsymbol{E}$ 表示 n 阶单位阵，则称 $\boldsymbol{B}_{\mathrm{J}}$ 为雅可比迭代矩阵.

7.2.2 高斯—塞德尔迭代法

由雅可比方法迭代公式(7.12)可知，在迭代的每一步计算过程中是用 $\boldsymbol{x}^{(k)}$ 的全部分量来计算 $\boldsymbol{x}^{(k+1)}$ 的所有分量，显然在计算第 i 个分量 $x_i^{(k+1)}$ 时，已经计算出的最新分量 $x_1^{(k+1)}, \cdots, x_{i-1}^{(k+1)}$ 没有被利用. 从直观上看，最新计算出的分量可能比旧的分量要好些. 因此，对这些最新计算出来的第 $k+1$ 次近似 $x^{(k+1)}$ 的分量 $x_j^{(k+1)}$ 加以利用，就得到所谓解方程组的高斯—塞德尔(Gauss－Seidel)迭代法(简称 G－S 方法)，其迭代格式的分量形式为

$$\begin{cases} x_1^{(k+1)} = \dfrac{1}{a_{11}}(b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)} - \cdots - a_{1n}x_n^{(k)}), \\ x_2^{(k+1)} = \dfrac{1}{a_{22}}(b_2 - a_{21}x_1^{(k+1)} - a_{23}x_3^{(k)} - \cdots - a_{2n}x_n^{(k)}), \\ \cdots\cdots \\ x_n^{(k+1)} = \dfrac{1}{a_{nn}}(b_n - a_{n1}x_1^{(k+1)} - a_{n2}x_2^{(k+1)} - \cdots - a_{n,n-1}x_{n-1}^{(k+1)}). \end{cases} \tag{7.15}$$

如果借助于求和符号，可以将式(7.15)写成更加简洁的形式

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\Big(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\Big) \quad (i = 1,2,\cdots,n). \tag{7.16}$$

利用式(7.10)及矩阵乘法，式(7.15)可以改写成如下的矩阵形式

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{D}^{-1}(\boldsymbol{L}\boldsymbol{x}^{(k+1)} + \boldsymbol{U}\boldsymbol{x}^{(k)}) + \boldsymbol{D}^{-1}\boldsymbol{b},$$

解出 $\boldsymbol{x}^{k+1}$，则有

$$\begin{aligned} \boldsymbol{x}^{(k+1)} &= (\boldsymbol{E}-\boldsymbol{D}^{-1}\boldsymbol{L})^{-1}\boldsymbol{D}^{-1}\boldsymbol{U}\boldsymbol{x}^{(k)} + (\boldsymbol{E}-\boldsymbol{D}^{-1}\boldsymbol{L})^{-1}\boldsymbol{D}^{-1}\boldsymbol{b} \\ &= (\boldsymbol{D}-\boldsymbol{L})^{-1}\boldsymbol{U}\boldsymbol{x}^{(k)} + (\boldsymbol{D}-\boldsymbol{L})^{-1}\boldsymbol{b}. \end{aligned} \tag{7.17}$$

记

$$\boldsymbol{B}_{\mathrm{G}} = (\boldsymbol{D}-\boldsymbol{L})^{-1}\boldsymbol{U}, \boldsymbol{f}_{\mathrm{G}} = (\boldsymbol{D}-\boldsymbol{L})^{-1}\boldsymbol{b},$$

则 $\boldsymbol{B}_{\mathrm{G}}$ 是高斯—塞德尔迭代法的迭代矩阵.

例 7.2 分别用雅可比迭代法和高斯—塞德尔迭代法求解方程组

$$\begin{cases} 10x_1 + 3x_2 + x_3 = 14, \\ 2x_1 - 10x_2 + 3x_3 = -5, \\ x_1 + 3x_2 + 10x_3 = 14. \end{cases}$$

设方程组的准确解为 $\boldsymbol{x}^*$，取初值 $\boldsymbol{x}^{(0)}=(0,0,0)^{\mathrm{T}}$，要求 $\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^*\|_\infty<9\times10^{-3}$.

解 从后面判断收敛性的知识可知，用这两种方法求解该方程组均收敛．方程组的精确解是 $\boldsymbol{x}^*=(1,1,1)^{\mathrm{T}}$.

(1)用雅可比方法计算．易见其雅可比迭代公式为

$$\begin{cases} x_1^{(k+1)}=(14-3x_2^{(k)}-x_3^{(k)})/10, \\ x_2^{(k+1)}=(-5-2x_1^{(k)}-3x_3^{(k)})/(-10), \\ x_3^{(k+1)}=(14-x_1^{(k)}-3x_2^{(k)})/10. \end{cases}$$

当取 $\boldsymbol{x}^{(0)}=(0,0,0)^{\mathrm{T}}$ 迭代 6 次后的计算值见表 7.1.

表 7.1 迭代计算结果

$\boldsymbol{x}^{(1)}$	$(1.4,0.5,1.4)^{\mathrm{T}}$
$\boldsymbol{x}^{(2)}$	$(1.11,1.20,1.11)^{\mathrm{T}}$
$\boldsymbol{x}^{(3)}$	$(0.929,1.055,0.929)^{\mathrm{T}}$
$\boldsymbol{x}^{(4)}$	$(0.9906,0.9405,0.9906)^{\mathrm{T}}$
$\boldsymbol{x}^{(5)}$	$(1.0116,0.9953,1.0116)^{\mathrm{T}}$
$\boldsymbol{x}^{(6)}$	$(1.00025,1.00580,1.00025)^{\mathrm{T}}$

(2)用高斯—塞德尔迭代法计算．易见高斯—塞德尔迭代公式为

$$\begin{cases} x_1^{(k+1)}=(14-3x_2^{(k)}-x_3^{(k)})/10, \\ x_2^{(k+1)}=(-5-2x_1^{(k+1)}-3x_3^{(k)})/(-10), \\ x_3^{(k+1)}=(14-x_1^{(k+1)}-3x_2^{(k+1)})/10, \end{cases}$$

取同样的初值，迭代四次后得到

$$\boldsymbol{x}^{(4)}=(0.99154,0.99578,1.0021)^{\mathrm{T}}.$$

从而可见，本题选用高斯—塞德尔迭代法比用雅可比迭代法收敛快，但是要注意结论并非总是如此，甚至有这样的方程组，雅可比方法收敛，而高斯—塞德尔迭代法却是发散的，在后面将举例说明．

7.2.3 逐次超松弛迭代法

为了加快收敛速度，将高斯—塞德尔迭代格式(7.16)改写成

$$x_i^{(k+1)}=x_i^{(k)}+\frac{1}{a_{ii}}\left(b_i-\sum_{j=1}^{i-1}a_{ij}x_j^{(k+1)}-\sum_{j=i}^{n}a_{ij}x_j^{(k)}\right)\quad(i=1,2,\cdots,n).$$

上式等号右端的第二项可以看成是在 $x_i^{(k)}$ 的基础上做的一个修正量．为了获得更快的收敛速度，在修正量前乘以一个常数 ω，称为松弛因子，从而得到所谓的逐次超松弛迭代法(Successive Over Relaxation Method，简称 SOR 方法)，即

$$x_i^{(k+1)}=x_i^{(k)}+\frac{\omega}{a_{ii}}\left(b_i-\sum_{j=1}^{i-1}a_{ij}x_j^{(k+1)}-\sum_{j=i}^{n}a_{ij}x_j^{(k)}\right)\quad(i=1,2,\cdots,n).\tag{7.18}$$

或者

$$\begin{cases} x_1^{(k+1)} = x_1^{(k)} + \dfrac{\omega}{a_{11}}(b_1 - a_{11}x_1^{(k)} - a_{12}x_2^{(k)} - a_{13}x_3^{(k)} - \cdots - a_{1n}x_n^{(k)}), \\ x_2^{(k+1)} = x_2^{(k)} + \dfrac{\omega}{a_{22}}(b_2 - a_{21}x_1^{(k+1)} - a_{22}x_2^{(k)} - a_{23}x_3^{(k)} - \cdots - a_{2n}x_n^{(k)}), \\ \cdots\cdots \\ x_n^{(k+1)} = x_n^{(k)} + \dfrac{\omega}{a_{nn}}(b_n - a_{n1}x_1^{(k+1)} - a_{n2}x_2^{(k+1)} - \cdots - a_{n,n-1}x_{n-1}^{(k+1)} - a_{nn}x_n^{(k)}). \end{cases} \tag{7.19}$$

同理,从迭代的分量形式(7.19)可以得到迭代的矩阵形式

$$\begin{aligned} \boldsymbol{x}^{(k+1)} &= \boldsymbol{x}^{(k)} + \omega\boldsymbol{D}^{-1}(\boldsymbol{b} + \boldsymbol{L}\boldsymbol{x}^{(k+1)} + \boldsymbol{U}\boldsymbol{x}^{(k)} - \boldsymbol{D}\boldsymbol{x}^{(k)}) \\ &= \omega\boldsymbol{D}^{-1}\boldsymbol{L}\boldsymbol{x}^{(k+1)} + [\boldsymbol{E} + \omega\boldsymbol{D}^{-1}(\boldsymbol{U} - \boldsymbol{D})]\boldsymbol{x}^{(k)} + \omega\boldsymbol{D}^{-1}\boldsymbol{b}. \end{aligned}$$

解出 $\boldsymbol{x}^{(k+1)}$ 得到

$$\begin{aligned} \boldsymbol{x}^{(k+1)} &= (\boldsymbol{E} - \omega\boldsymbol{D}^{-1}\boldsymbol{L})^{-1}[\boldsymbol{E} + \omega\boldsymbol{D}^{-1}(\boldsymbol{U} - \boldsymbol{D})]\boldsymbol{x}^{(k)} + (\boldsymbol{E} - \omega\boldsymbol{D}^{-1}\boldsymbol{L})^{-1}\omega\boldsymbol{D}^{-1}\boldsymbol{b} \\ &= (\boldsymbol{D} - \omega\boldsymbol{L})^{-1}[(1-\omega)\boldsymbol{D} + \omega\boldsymbol{U}]\boldsymbol{x}^{(k)} + \omega(\boldsymbol{D} - \omega\boldsymbol{L})^{-1}\boldsymbol{b}. \end{aligned} \tag{7.20}$$

记

$$\boldsymbol{B}_S = (\boldsymbol{D} - \omega\boldsymbol{L})^{-1}[(1-\omega)\boldsymbol{D} + \omega\boldsymbol{U}], \boldsymbol{f}_S = \omega(\boldsymbol{D} - \omega\boldsymbol{L})^{-1}\boldsymbol{b}$$

称 $\boldsymbol{B}_S$ 为 SOR 方法的迭代矩阵.

注:

(1)当 $\omega=1$ 时,SOR 方法就是 G—S 方法.

(2)SOR 方法是 G—S 方法的一种修正,可由下述思想得到. 设已知 $\boldsymbol{x}^{(k)}$ 及已计算 $\boldsymbol{x}^{(k+1)}$ 的分量 $x_j^{(k+1)}(j=1,2,\cdots,i-1)$.

① 首先由 G—S 迭代法定义一个辅助量 $\tilde{x}_i^{(k+1)}$

$$\tilde{x}_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right) \quad (i = 1,2,\cdots,n). \tag{7.21}$$

② 再由 $x_i^{(k)}$ 和 $\tilde{x}_i^{(k+1)}$ 加权平均得到 $x_i^{(k+1)}$,即

$$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \omega\tilde{x}_i^{(k+1)} = x_i^{(k)} + \omega(\tilde{x}_i^{(k+1)} - x_i^{(k)}). \tag{7.22}$$

将式(7.21)代入式(7.22)就得到式(7.18).

(3)SOR 法的收敛性和收敛的快慢与松弛因子 ω 有密切关系.

例 7.3 用 SOR 方法求解线性代数方程组

$$\begin{cases} -4x_1 + x_2 + x_3 + x_4 = 1, \\ x_1 - 4x_2 + x_3 + x_4 = 1, \\ x_1 + x_2 - 4x_3 + x_4 = 1, \\ x_1 + x_2 + x_3 - 4x_4 = 1. \end{cases}$$

初始值取为 $\boldsymbol{x}^{(0)}=(0,0,0,0)^{\mathrm{T}}$，其精确解为$(-1,-1,-1,-1)^{\mathrm{T}}$，要求精确到 10^{-5}.

解 SOR 法的迭代公式为

$$\begin{cases} x_1^{(k+1)} = x_1^{(k)} - \dfrac{\omega}{4}(1+4x_1^{(k)} - x_2^{(k)} - x_3^{(k)} - x_4^{(k)}), \\ x_2^{(k+1)} = x_2^{(k)} - \dfrac{\omega}{4}(1-x_1^{(k+1)} + 4x_2^{(k)} - x_3^{(k)} - x_4^{(k)}), \\ x_3^{(k+1)} = x_3^{(k)} - \dfrac{\omega}{4}(1-x_1^{(k+1)} - x_2^{(k+1)} + 4x_3^{(k)} - x_4^{(k)}), \\ x_4^{(k+1)} = x_4^{(k)} - \dfrac{\omega}{4}(1-x_1^{(k+1)} - x_2^{(k+1)} - x_3^{(k+1)} - 4x_4^{(k)}). \end{cases} \quad (k=0,1,2,\cdots)$$

当 ω 取不同的值时，迭代加速的效果是不一样的，下面分别举出一些不同 ω 取值的计算结果.

取 $\omega=1.0$ 时，迭代 21 次，得到

$$\boldsymbol{x}^{(21)} = (-0.99999, -0.99999, -1.00000, -1.00000)^{\mathrm{T}};$$

取 $\omega=1.1$ 时，迭代 16 次，得到

$$\boldsymbol{x}^{(16)} = (-0.99999, -1.00000, -1.00000, -1.00000)^{\mathrm{T}};$$

取 $\omega=1.2$ 时，迭代 11 次，得到

$$\boldsymbol{x}^{(11)} = (-1.00000, -1.00000, -1.00000, -1.00000)^{\mathrm{T}};$$

取 $\omega=1.3$ 时，迭代 10 次，得到

$$\boldsymbol{x}^{(10)} = (-1.00000, -1.00000, -1.00000, -1.00000)^{\mathrm{T}};$$

取 $\omega=1.4$ 时，迭代 13 次，得到

$$\boldsymbol{x}^{(13)} = (-1.00000, -1.00000, -1.00000, -1.00000)^{\mathrm{T}};$$

取 $\omega=1.5$ 时，迭代 16 次，得到

$$\boldsymbol{x}^{(16)} = (-0.99999, -1.00000, -1.00000, -1.00000)^{\mathrm{T}};$$

取 $\omega=1.6$ 时，迭代 22 次，得到

$$\boldsymbol{x}^{(22)} = (-1.00000, -1.00000, -1.00000, -1.00000)^{\mathrm{T}};$$

取 $\omega=1.7$ 时，迭代 32 次，得到

$$\boldsymbol{x}^{(11)} = (-1.00000, -0.99999, -1.00000, -1.00000)^{\mathrm{T}};$$

取 $\omega=1.8$ 时，迭代 52 次，得到

$$\boldsymbol{x}^{(52)} = (-0.99999, -1.00000, -1.00000, -0.99999)^{\mathrm{T}}.$$

显然，松弛因子选得好，会使 SOR 法的收敛速度大大加快. 本题选 $\omega=1.3$，加速效果最明显. 一般而言，要得到一个最佳的松弛因子是有难度的.

对于一个迭代法而言，不可能无限迭代下去. 总是希望迭代到某一步后就停止迭代，这时

的解就非常接近于精确解，这就涉及收敛性和收敛速度的问题．下一节主要介绍迭代法收敛性的一些判定定理，最后简单提及收敛速度的定义．

7.3 迭代法的收敛性

7.3.1 一阶定常迭代的基本定理

定义 7.1 设有矩阵序列 $\boldsymbol{A}_k=(a_{ij}^{(k)})_{n\times n}(k=1,2,\cdots)$ 及 $\boldsymbol{A}=(a_{ij})_{n\times n}$，如果 n^2 个数列的极限存在且有

$$\lim_{k\to\infty}a_{ij}^{(k)}=a_{ij}\quad(i,j=1,2,\cdots,n)$$

成立，则称$\{\boldsymbol{A}_k\}$收敛于 $\boldsymbol{A}$，记为$\lim\limits_{k\to\infty}\boldsymbol{A}_k=\boldsymbol{A}$.

例 7.4 矩阵序列

$$\boldsymbol{A}=\begin{bmatrix}\lambda & 1\\ 0 & \lambda\end{bmatrix},\boldsymbol{A}^2=\begin{bmatrix}\lambda^2 & 2\lambda\\ 0 & \lambda^2\end{bmatrix},\cdots,\boldsymbol{A}^k=\begin{bmatrix}\lambda^k & k\lambda^{k-1}\\ 0 & \lambda^k\end{bmatrix},\cdots$$

当$|\lambda|<1$时，$\boldsymbol{A}^k\to\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}$(当 $k\to\infty$时).

矩阵序列极限的概念可以用任何矩阵范数来描述．

定理 7.1 $\lim\limits_{k\to\infty}\boldsymbol{A}_k=\boldsymbol{A}$ 的充要条件是 $\|\boldsymbol{A}_k-\boldsymbol{A}\|\to 0\quad(k\to\infty)$.

定理 7.2 $\lim\limits_{k\to\infty}\boldsymbol{A}_k=\boldsymbol{A}\Leftrightarrow$对任意向量 $\boldsymbol{x}\in\boldsymbol{R}^n$ 都有$\lim\limits_{k\to\infty}\boldsymbol{A}_k\boldsymbol{x}=\boldsymbol{A}\boldsymbol{x}$

在 7.1 节的最后讲到的判断迭代是否收敛就转化为矩阵 $\boldsymbol{B}^k$ 是否收敛于零矩阵的问题．

定理 7.3 设 $\boldsymbol{B}=(b_{ij})_{n\times n}$，则 $\boldsymbol{B}^k\to\boldsymbol{O}(k\to\infty)$的充要条件是 $\rho(\boldsymbol{B})<1$.

定理 7.4 (迭代法基本定理)设有方程组

$$\boldsymbol{x}=\boldsymbol{B}\boldsymbol{x}+\boldsymbol{f},\tag{7.23}$$

对于任意初始向量 $\boldsymbol{x}^{(0)}$，解此方程组的迭代法(即 $\boldsymbol{x}^{(k+1)}=\boldsymbol{B}\boldsymbol{x}^{(k)}+\boldsymbol{f}$)收敛的充要条件是 $\rho(\boldsymbol{B})<1$.

证明 充分性．设 $\rho(\boldsymbol{B})<1$. 从而 1 不是 $\boldsymbol{B}$ 的特征值，故

$$|1\cdot\boldsymbol{E}-\boldsymbol{B}|=|\boldsymbol{E}-\boldsymbol{B}|\neq 0,$$

即 $\boldsymbol{E}-\boldsymbol{B}$ 是可逆阵，从而方程组$(\boldsymbol{E}-\boldsymbol{B})\boldsymbol{x}=\boldsymbol{f}$ 有唯一解，即 $\boldsymbol{x}=\boldsymbol{B}\boldsymbol{x}+\boldsymbol{f}$ 有唯一解，记为 $\boldsymbol{x}^*$，即

$$\boldsymbol{x}^*=\boldsymbol{B}\boldsymbol{x}^*+\boldsymbol{f}.\tag{7.24}$$

误差向量

$$\boldsymbol{\varepsilon}^{(k)}=\boldsymbol{x}^{(k)}-\boldsymbol{x}^*=\boldsymbol{B}^k\boldsymbol{\varepsilon}^{(0)},\boldsymbol{\varepsilon}^{(0)}=\boldsymbol{x}^{(0)}-\boldsymbol{x}^*,$$

由设 $\rho(\boldsymbol{B})<1$，应用定理 7.3，有 $\boldsymbol{B}^k\to\boldsymbol{O}(k\to\infty)$. 于是对任意 $\boldsymbol{x}^{(0)}$，有 $\boldsymbol{\varepsilon}^k\to 0(k\to\infty)$即 $\boldsymbol{x}^{(k)}\to\boldsymbol{x}^*$ $(k\to\infty)$.

必要性．设对任意 $\boldsymbol{x}^{(0)}$ 皆有

$$\lim_{(k\to\infty)} \boldsymbol{x}^{(k)} = \boldsymbol{x}^*,$$

其中 $\boldsymbol{x}^{(k+1)}=\boldsymbol{B}\boldsymbol{x}^{(k)}+\boldsymbol{f}$. 显然,极限 $\boldsymbol{x}^*$ 是方程组式(7.23)的解,且对任意 $\boldsymbol{x}^{(0)}$ 有

$$\boldsymbol{\varepsilon}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}^* = \boldsymbol{B}^k\boldsymbol{\varepsilon}^{(0)} \to 0 \quad (k\to\infty).$$

由定理 7.2 知

$$\boldsymbol{B}^{(k)} \to \boldsymbol{O} \quad (k\to\infty).$$

再由定理 7.3,即得 $\rho(\boldsymbol{B})<1$.

推论 设 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$ 且 $\boldsymbol{A}=\boldsymbol{D}-\boldsymbol{L}-\boldsymbol{U}$,$\boldsymbol{A}$,$\boldsymbol{D}$ 均为非奇异阵,则

(1)解方程组的雅可比迭代法收敛的充要条件是 $\rho(\boldsymbol{B}_J)<1$,其中

$$\boldsymbol{B}_J = \boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{U}) = \boldsymbol{E}-\boldsymbol{D}^{-1}\boldsymbol{A}.$$

(2)解方程组的高斯—塞德尔迭代法收敛的充要条件是 $\rho(\boldsymbol{B}_G)<1$,其中

$$\boldsymbol{B}_G = (\boldsymbol{D}-\boldsymbol{L})^{-1}\boldsymbol{U}.$$

(3)解方程组的 SOR 方法收敛的充要条件是 $\rho(\boldsymbol{B}_S)<1$,其中

$$\boldsymbol{B}_S = (\boldsymbol{D}-\omega\boldsymbol{L})^{-1}[(1-\omega)\boldsymbol{D}+\omega\boldsymbol{U}].$$

定理 7.4 是判断一阶定常迭代法收敛的基本理论.

例 7.5 考察用迭代法解下述方程组的收敛性:

$$\boldsymbol{x} = \boldsymbol{B}\boldsymbol{x}+\boldsymbol{f},$$

其中,$\boldsymbol{B}=\begin{bmatrix}0 & 2\\ 3 & 0\end{bmatrix}$,$\boldsymbol{f}=\begin{bmatrix}5\\ 5\end{bmatrix}$.

解 矩阵 $\boldsymbol{B}$ 的特征方程为

$$\det(\lambda\boldsymbol{E}-\boldsymbol{B}) = \lambda^2-6=0.$$

特征根 $\lambda_{1,2}=\pm\sqrt{6}$,即 $\rho(\boldsymbol{B})>1$. 这说明用迭代法解此方程组不收敛.

例 7.6 设有方程组

$$\begin{bmatrix}1 & 2 & -2\\ 1 & 1 & 1\\ 2 & 2 & 1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix} = \begin{bmatrix}b_1\\ b_2\\ b_3\end{bmatrix},$$

试考察用雅可比迭代法和高斯—塞德尔迭代法求解的收敛性.

解 雅可比迭代法的迭代矩阵

$$\boldsymbol{B}_J = \boldsymbol{E}-\boldsymbol{D}^{-1}\boldsymbol{A} = \begin{bmatrix}0 & -2 & 2\\ -1 & 0 & -1\\ -2 & -2 & 0\end{bmatrix},$$

其特征值为 $\lambda_{1,2,3}=0$,故谱半径 $\rho(\boldsymbol{B}_J)=0<1$,从而由定理 7.4 可知雅可比迭代法收敛.

高斯—塞德尔迭代法的迭代矩阵为

$$\boldsymbol{B}_{\mathrm{G}}=(\boldsymbol{D}-\boldsymbol{L})^{-1}\boldsymbol{U}=\begin{bmatrix}0 & -2 & 2\\ 0 & 2 & -3\\ 0 & 0 & 2\end{bmatrix}.$$

其特征值为 $\lambda_1=0,\lambda_{2,3}=2$，故谱半径为 $\rho(\boldsymbol{B}_{\mathrm{G}})=2>1$，所以高斯—塞德尔迭代法发散.

例 7.7 给定线性代数方程组 $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$，其中

$$\boldsymbol{A}=\begin{bmatrix}1 & 0.5 & 0.5\\ 0.5 & 1 & 0.5\\ 0.5 & 0.5 & 1\end{bmatrix},$$

试考察用雅可比迭代法和高斯—塞德尔迭代法求解的收敛性.

解 雅可比迭代的迭代矩阵为

$$\boldsymbol{B}_{\mathrm{J}}=\boldsymbol{E}-\boldsymbol{D}^{-1}\boldsymbol{A}=\begin{bmatrix}0 & -0.5 & -0.5\\ -0.5 & 0 & -0.5\\ -0.5 & -0.5 & 0\end{bmatrix}$$

其特征值为 $\lambda_1=-1,\lambda_{2,3}=0.5$ 从而谱半径 $\rho(\boldsymbol{B}_{\mathrm{J}})=1$，雅可比迭代法发散.

注意到矩阵 $\boldsymbol{A}$ 为对称阵且各阶顺序主子式

$$|1|=1>0,\begin{vmatrix}1 & 0.5\\ 0.5 & 1\end{vmatrix}=0.75>0,|\boldsymbol{A}|=0.5>0,$$

从而矩阵 $\boldsymbol{A}$ 为对称正定阵，由定理 7.10 的推论可知高斯—塞德尔迭代法收敛.

注：(1)从文中叙述 G—S 迭代法的过程中不难得到，G—S 迭代法似乎比雅可比迭代法要好些，从例 7.2 中还可以看出，G—S 迭代法比雅可比迭代法收敛快些. 现在从例 7.6 和例 7.7 中可以看出，有的问题两者的收敛性可能完全相反. 也就是说完全可能出现一种迭代法收敛而另一种却发散的情况.

(2)定理 7.4 只是说 $\rho(\boldsymbol{B})<1$ 是对于任意选取的初始迭代向量迭代法都收敛的充要条件，并没有说当 $\rho(\boldsymbol{B})\geqslant 1$ 时，对于任意选取的初始迭代向量，迭代法一定都要发散. 实际上，在例 7.6 和例 7.7 中所说的“发散”，其准确含义应该理解为“不是对任意初始向量 $\boldsymbol{x}^{(0)}$ 都收敛”，也就是说“对有的初始向量迭代法发散，对有的初始迭代向量它可能收敛”. 下面的例子正好可以说明这一点.

例 7.8 设 n 阶矩阵 $\boldsymbol{B}$ 的谱半径 $\rho(\boldsymbol{B})\geqslant 1$，但是 $\boldsymbol{B}$ 有一个特征值 λ，其模 $|\lambda|<1$，则一定存在初始向量 $\boldsymbol{x}^{(0)}$，使得迭代法

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{B}\boldsymbol{x}^{(k)}+\boldsymbol{f}\quad(k=0,1,2,\cdots)$$

关于此初始迭代向量收敛到 $\boldsymbol{x}=\boldsymbol{B}\boldsymbol{x}+\boldsymbol{f}$ 的准确解 x^*.

证明 设矩阵 $\boldsymbol{B}$ 相应于特征值 λ 的特征向量为 $\boldsymbol{y}$，则有

$$\boldsymbol{B}\boldsymbol{y}=\lambda\boldsymbol{y},\boldsymbol{B}^k\boldsymbol{y}=\lambda^k\boldsymbol{y}$$

由式(7.7)和式(7.8)可知

$$\boldsymbol{x}^{(k)}-\boldsymbol{x}^{*}=\boldsymbol{B}^{k}(\boldsymbol{x}^{(0)}-\boldsymbol{x}^{*}) \tag{7.25}$$

取初始向量

$$\boldsymbol{x}^{(0)}=\boldsymbol{x}^{*}+\boldsymbol{y}$$

则有

$$\boldsymbol{x}^{(k)}-\boldsymbol{x}^{*}=\boldsymbol{B}^{k}(\boldsymbol{x}^{(0)}-\boldsymbol{x}^{*})=\boldsymbol{B}^{k}\boldsymbol{y}=\lambda^{k}\boldsymbol{y}$$

注意到$|\lambda|<1$,从而

$$\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{*}\|=|\lambda|^{k}\|\boldsymbol{y}\|\to 0$$

即$\boldsymbol{x}^{(k)}\to\boldsymbol{x}^{*}$.

运用定理 7.4 判断收敛性需要计算迭代矩阵的特征值,由定理 6.5 可知$\rho(\boldsymbol{B})\leqslant\|\boldsymbol{B}\|$,即矩阵的谱半径小于等于矩阵的任何一种算子范数,当$\|\boldsymbol{B}\|<1$时,还可以给出如下的充分条件.

定理 7.5 (迭代法法收敛的充分条件)设有方程组

$$\boldsymbol{x}=\boldsymbol{B}\boldsymbol{x}+\boldsymbol{f},\boldsymbol{B}\in\boldsymbol{R}^{n\times n}.$$

记一阶定常迭代法

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{B}\boldsymbol{x}^{(k)}+\boldsymbol{f}$$

如果迭代矩阵的某一种范数$\|\boldsymbol{B}\|=q<1$,则

(1)迭代法收敛,即对任意初始迭代向量$\boldsymbol{x}^{(0)}$,有

$\lim\limits_{k\to\infty}\boldsymbol{x}^{(k)}=\boldsymbol{x}^{*}$且$\boldsymbol{x}^{*}=\boldsymbol{B}\boldsymbol{x}^{*}+\boldsymbol{f}$;

(2)$\|\boldsymbol{x}^{*}-\boldsymbol{x}^{(k)}\|\leqslant q^{k}\|\boldsymbol{x}^{*}-\boldsymbol{x}^{(0)}\|$;

(3)$\|\boldsymbol{x}^{*}-\boldsymbol{x}^{(k)}\|\leqslant\dfrac{q}{1-q}\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\|$;

(4)$\|\boldsymbol{x}^{*}-\boldsymbol{x}^{(k)}\|\leqslant\dfrac{q^{k}}{1-q}\|\boldsymbol{x}^{(1)}-\boldsymbol{x}^{(0)}\|$.

证明 (1)利用定理 7.4,结论(1)是显然的.

(2)注意到

$$\boldsymbol{x}^{*}-\boldsymbol{x}^{(k+1)}=\boldsymbol{B}(\boldsymbol{x}^{*}-\boldsymbol{x}^{(k)}),$$

以及

$$\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)}=\boldsymbol{B}(\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}),$$

于是有

① $\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)}\|\leqslant q\|\boldsymbol{x}^{(k)}-\boldsymbol{x}^{(k-1)}\|\quad(k=1,2,\cdots)$;

② $\|\boldsymbol{x}^{*}-\boldsymbol{x}^{(k+1)}\|\leqslant q\|\boldsymbol{x}^{*}-\boldsymbol{x}^{(k)}\|$.

反复利用②就得到(2).

(3)注意到

$$\begin{aligned}\| \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)} \| &= \| \boldsymbol{x}^* - \boldsymbol{x}^{(k)} - (\boldsymbol{x}^* - \boldsymbol{x}^{(k+1)}) \| \\ &\geqslant \| \boldsymbol{x}^* - \boldsymbol{x}^{(k)} \| - \| \boldsymbol{x}^* - \boldsymbol{x}^{(k+1)} \| \\ &\geqslant (1-q) \| \boldsymbol{x}^* - \boldsymbol{x}^{(k)} \| ,\end{aligned}$$

从而

$$\begin{aligned}\| \boldsymbol{x}^* - \boldsymbol{x}^{(k)} \| &\leqslant \frac{1}{1-q} \| \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)} \| \\ &\leqslant \frac{q}{1-q} \| \boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)} \| \quad (k = 1,2,\cdots).\end{aligned}$$

(4)在结论(3)的基础上反复利用①即得.

要特别注意,当矩阵 $\boldsymbol{B}$ 的某种算子范数 $\| \boldsymbol{B} \| > 1$ 时,并不能判断迭代法发散,例如

$$\boldsymbol{B} = \begin{pmatrix} 0.9 & 0 \\ 0.2 & 0.8 \end{pmatrix},$$

易见 $\| \boldsymbol{B} \|_\infty = 1.0$, $\| \boldsymbol{B} \|_1 = 1.1$,

虽然 $\boldsymbol{B}$ 的这些范数都大于 1,但 $\boldsymbol{B}$ 的特征值为 $\lambda_1 = 0.9, \lambda_2 = 0.8$,由定理 7.4,对此方程组应用迭代法还是收敛的.

由定理 7.5 可知,$\| \boldsymbol{B} \| = q < 1$ 越小,由结论(2)知道迭代法收敛越快. 由结论(3)可知,当 $\boldsymbol{B}$ 的某一种范数 $\| \boldsymbol{B} \| < 1$ 时,若相邻两次迭代误差 $\| \boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)} \| < \varepsilon_0$($\varepsilon_0$ 为给定的精度要求),则可以说明第 k 次迭代值和准确值 $\boldsymbol{x}^*$ 之间的误差就充分小. 所以实际中可用 $\| \boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)} \| < \varepsilon_0$ 作为判断迭代终止的条件. 另外结论(4)是一种先验误差估计,可以用来事先确定迭代多少次数才能保证精度.

7.3.2 某些特殊方程组迭代法的收敛性

在科学及工程计算中,某些方程组的系数矩阵 $\boldsymbol{A}$ 常常具有某种特殊性比如对角占优或对称正定等性质,下面就讨论具有这些特殊系数矩阵的迭代法的收敛性.

定义 7.3 设 $\boldsymbol{A} = (a_{ij})_{n\times n}$,如果矩阵 $\boldsymbol{A}$ 满足条件

$$| a_{ii} | > \sum_{\substack{j=1 \\ j\neq i}}^{n} | a_{ij} | \quad (i = 1,2,\cdots,n), \tag{7.26}$$

即 $\boldsymbol{A}$ 的每行对角元素的绝对值都严格大于同行其他元素绝对值之和,则称 $\boldsymbol{A}$ 为(行)严格对角优势矩阵.

例 7.9 $\boldsymbol{A} = \begin{bmatrix} -4 & 1 & 0 & 0 \\ 1 & -4 & 1 & 0 \\ 0 & 1 & -4 & 1 \\ 0 & 0 & 1 & -4 \end{bmatrix}$ 为严格对角优势矩阵.

定理 7.6 (对角占优定理)如果 $\boldsymbol{A} = (a_{ij})_{n\times n} \in R^{n\times n}$ 为严格对角优势阵,则 $\boldsymbol{A}$ 是非奇异矩阵.

证明 设 $\boldsymbol{A}$ 为严格对角优势阵,采用反证法. 若 $|\boldsymbol{A}| = 0$,则 $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$ 有非零解,记为 $\boldsymbol{x} =$

$(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$. 又记$|x_k|=\max\limits_{1\leqslant i\leqslant n}|x_i|\neq 0$,由齐次方程组的第 k 个方程

$$\sum_{j=1}^{n}a_{kj}x_j=0,$$

则有

$$|a_{kk}x_k|=\Big|\sum_{\substack{j=1\\j\neq k}}^{n}a_{kj}x_j\Big|\leqslant\sum_{\substack{j=1\\j\neq k}}^{n}|a_{kj}||x_j|\leqslant|x_k|\sum_{\substack{j=1\\j\neq k}}^{n}|a_{kj}|,$$

从而得到

$$|a_{kk}|\leqslant\sum_{\substack{j=1\\j\neq k}}^{n}|a_{kj}|$$

与假设矛盾,故$|\boldsymbol{A}|\neq 0$

定理 7.7 如果$\boldsymbol{A}\in\boldsymbol{R}^{n\times n}$为严格对角优势矩阵,则对任意的 $\boldsymbol{x}^{(0)}$,解方程组 $\boldsymbol{Ax}=\boldsymbol{b}$ 的雅可比迭代法,高斯—塞德尔迭代法均收敛.

证明 设 $\boldsymbol{A}$ 为(行)严格对角占优阵.

(1)由假设知 $a_{ii}\neq 0$,雅可比迭代矩阵 $\boldsymbol{B}=\boldsymbol{E}-\boldsymbol{D}^{-1}\boldsymbol{A}=(b_{ij})_{n\times n}$,其中 $b_{ij}=-a_{ij}/a_{ii}$,$j\neq i$,$b_{ii}=0$. 从而

$$\|\boldsymbol{B}\|_{\infty}=\max_{1\leqslant i\leqslant n}\sum_{j=1}^{n}|b_{ij}|=\max_{1\leqslant i\leqslant n}\frac{\sum\limits_{j\neq i}|a_{ij}|}{|a_{ii}|}<1.$$

由定理 7.5 可知雅可比迭代法收敛. 该结果也可以按照下面证明 G—S 迭代法收敛的方法类似证明.

(2)高斯—塞德尔迭代法的迭代矩阵为

$$\boldsymbol{B}_{\mathrm{G}}=(\boldsymbol{D}-\boldsymbol{L})^{-1}\boldsymbol{U}.$$

又 $$|\lambda\boldsymbol{E}-\boldsymbol{B}_{\mathrm{G}}|=|\lambda\boldsymbol{E}-(\boldsymbol{D}-\boldsymbol{L})^{-1}\boldsymbol{U}|=|(\boldsymbol{D}-\boldsymbol{L})^{-1}||\lambda(\boldsymbol{D}-\boldsymbol{L})-\boldsymbol{U}|=0,$$

注意到$|(\boldsymbol{D}-\boldsymbol{L})^{-1}|\neq 0$,从而 $\boldsymbol{B}_{\mathrm{G}}$ 的特征值满足

$$|\lambda(\boldsymbol{D}-\boldsymbol{L})-\boldsymbol{U}|=0.$$

记

$$\boldsymbol{C}=\lambda(\boldsymbol{D}-\boldsymbol{L})-\boldsymbol{U}=\begin{bmatrix}\lambda a_{11} & a_{12} & a_{13} & \cdots & a_{1n}\\ \lambda a_{21} & a_{22} & a_{23} & \cdots & a_{2n}\\ \cdots & \cdots & \cdots & & \cdots\\ \lambda a_{n1} & \lambda a_{n2} & \lambda a_{n3} & \cdots & \lambda a_{nn}\end{bmatrix}.$$

下面来说明当$|\lambda|\geqslant 1$时$|\boldsymbol{C}|\neq 0$.

$$|c_{ii}|=|\lambda||a_{ii}|>|\lambda|\Big(\sum_{j=1}^{i-1}|a_{ij}|+\sum_{j=i+1}^{n}|a_{ij}|\Big)$$

$$\geqslant \sum_{j=1}^{i-1} |\lambda a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}| = \sum_{\substack{j=1 \\ j\neq i}}^{n} |c_{ij}|,$$

这表明当$|\lambda|\geqslant 1$时，矩阵$\boldsymbol{C}=\lambda(\boldsymbol{D}-\boldsymbol{L})-\boldsymbol{U}$为(行)严格对角占优阵，从而由定理7.6可知

$$|\boldsymbol{C}| = |\lambda(\boldsymbol{D}-\boldsymbol{L})-\boldsymbol{U}| \neq 0,$$

由此说明$\boldsymbol{B}_{\mathrm{G}}$的特征值均满足$|\lambda|<1$，从而G－S迭代法收敛.

下面讨论SOR迭代法迭代的收敛性.

定理7.8 (SOR收敛的必要条件)设解方程组$\boldsymbol{Ax}=\boldsymbol{b}$的迭代法收敛，则$0<\omega<2$.

证明 SOR法的迭代矩阵为$\boldsymbol{B}_{\mathrm{S}}=(\boldsymbol{D}-\omega\boldsymbol{L})^{-1}[(1-\omega)\boldsymbol{D}+\omega\boldsymbol{U}]$，设$\lambda_1,\lambda_2,\cdots,\lambda_n$是$\boldsymbol{B}_{\mathrm{S}}$的$n$个特征值，则(为不引起混淆，将矩阵$\boldsymbol{B}_{\mathrm{S}}$的行列式记为$\det \boldsymbol{B}_{\mathrm{S}}$)

$$\det \boldsymbol{B}_{\mathrm{S}} = \lambda_1\lambda_2\cdots\lambda_n,$$

$$|\det \boldsymbol{B}_{\mathrm{S}}| = |\lambda_1\lambda_2\cdots\lambda_n| \leqslant (\rho(\boldsymbol{B}_{\mathrm{S}}))^n, \tag{7.27}$$

而
$$\det \boldsymbol{B}_{\mathrm{S}} = \det(\boldsymbol{D}-\omega\boldsymbol{L})^{-1} \cdot \det[(1-\omega)\boldsymbol{D}+\omega\boldsymbol{U}].$$

注意到$\boldsymbol{D}-\omega\boldsymbol{L}$及$(1-\omega)\boldsymbol{D}+\omega\boldsymbol{U}$的表达式可知

$$\det(\boldsymbol{D}-\omega\boldsymbol{L})^{-1} = a_{11}^{-1}a_{22}^{-1}\cdots a_{nn}^{-1},$$

$$\det[(1-\omega)\boldsymbol{D}+\omega\boldsymbol{U}] = (1-\omega)a_{11}(1-\omega)a_{22}\cdots(1-\omega),$$

$$\begin{aligned} |\det \boldsymbol{B}_{\mathrm{S}}| &= |a_{11}^{-1}a_{22}^{-1}\cdots a_{nn}^{-1}| \cdot |(1-\omega)a_{11}(1-\omega)a_{22}\cdots(1-\omega)a_{nn}| \\ &= |1-\omega|^n. \end{aligned} \tag{7.28}$$

SOR法收敛，则$\rho(\boldsymbol{B}_{\mathrm{S}})<1$，从而由式(7.27)和式(7.28)可知$|1-\omega|<1$，从而结论得证.

该定理说明解$\boldsymbol{Ax}=\boldsymbol{b}$的SOR迭代法，只有当松弛因子满足$0<\omega<2$，才可能收敛.

定理7.9 设$\boldsymbol{Ax}=\boldsymbol{b}$，如果$\boldsymbol{A}$为严格对角占优且$0<\omega\leqslant 1$，则解$\boldsymbol{Ax}=\boldsymbol{b}$的SOR迭代法收敛.

定理7.10 设$\boldsymbol{Ax}=\boldsymbol{b}$，如果

(1)$\boldsymbol{A}$为对称正定矩阵；

(2)$0<\omega<2$；

则解$\boldsymbol{Ax}=\boldsymbol{b}$的SOR迭代法收敛.

注意到$\omega=1$时，SOR法就是G－S迭代法，从而得到

推论 若$\boldsymbol{A}$为对称正定阵，则解$\boldsymbol{Ax}=\boldsymbol{b}$的G－S迭代法收敛.

对于雅可比迭代法而言，有如下的结果.

定理7.11 若$\boldsymbol{A}$是对称正定阵，则雅可比迭代法收敛的充要条件是$2\boldsymbol{D}-\boldsymbol{A}$也对称正定.

下面简单介绍一下讨论迭代法的收敛速度. 可以证明$\rho(\boldsymbol{B})<1$且$\rho(\boldsymbol{B})$越小，则迭代法收敛越快. 一般地，对于方程组$\boldsymbol{x}=\boldsymbol{Bx}+\boldsymbol{f}$，定义如下：

定义7.4 称$R(\boldsymbol{B})=-\ln\rho(\boldsymbol{B})$为迭代法的收敛速度.

例7.10 讨论松弛因子$\omega=1.25$时，用SOR法求解方程组

$$\begin{cases} 4x_1+3x_2 = 16, \\ 3x_1+4x_2-x_3 = 20, \\ -x_2+4x_3 = -12, \end{cases}$$

的收敛性．若收敛，取初值 $\boldsymbol{x}^{(0)}=(0,0,0)^{\mathrm{T}}$ 迭代求解，使 $\|\boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)}\|_{\infty}<\frac{1}{2}\times 10^{-4}$．

解 方程组的系数矩阵是

$$\boldsymbol{A}=\begin{pmatrix}4 & 3 & 0\\ 3 & 4 & -1\\ 0 & -1 & 4\end{pmatrix}.$$

$\boldsymbol{A}$ 显然对称，又 $\boldsymbol{A}$ 的各阶顺序主子式

$$|4|>0,\quad \begin{vmatrix}4 & 3\\ 3 & 4\end{vmatrix}=7>0,\quad \begin{vmatrix}4 & 3 & 0\\ 3 & 4 & -1\\ 0 & -1 & 4\end{vmatrix}=31>0,$$

从而 $\boldsymbol{A}$ 对称正定，所以由定理 7.10 知 $\omega=1.25$ 时，SOR 法对任意初始迭代向量都收敛．迭代公式为

$$\begin{cases}x_1^{(k+1)}=x_1^{(k)}+\dfrac{1.25}{4}(16-4x_1^{(k)}-3x_2^{(k)}),\\ x_2^{(k+1)}=x_2^{(k)}+\dfrac{1.25}{4}(20-3x_1^{(k+1)}-4x_2^{(k)}+x_3^{(k)}),\\ x_3^{(k+1)}=x_3^{(k)}+\dfrac{1.25}{4}(-12+x_2^{(k+1)}-4x_3^{(k)}).\end{cases}$$

当 $\boldsymbol{x}^{(0)}=(0,0,0)^{\mathrm{T}}$ 时，计算可得

$$\begin{aligned}\boldsymbol{x}^{(1)}&=(5.00000,1.56250,-3.26172)^{\mathrm{T}},\\ \boldsymbol{x}^{(2)}&=(2.28516,2.69775,-2.0912)^{\mathrm{T}},\\ \boldsymbol{x}^{(3)}&=(1.89957,3.1412,-2.38415)^{\mathrm{T}},\\ &\cdots\cdots\\ \boldsymbol{x}^{(11)}&=(1.50005,3.33331,-2.166667)^{\mathrm{T}},\\ \boldsymbol{x}^{(12)}&=(1.50001,3.33333,-2.166667)^{\mathrm{T}}.\end{aligned}$$

由于 $\|\boldsymbol{x}^{(12)}-\boldsymbol{x}^{(11)}\|_{\infty}=0.00004<\frac{1}{2}\times 10^{-4}$，故方程组的近似解为

$$\boldsymbol{x}^{(12)}=(1.50001,3.33333,-2.166667)^{\mathrm{T}}.$$

习　题

1. 用迭代法解下列线性代数方程组

$$\begin{cases}10x_1+4x_2+4x_3=14,\\ 4x_1+10x_2+8x_3=12,\\ 4x_1+8x_2+10x_3=14.\end{cases}$$

(1)分别写出 Jacobi 迭代、G－S 迭代和 SOR 迭代($\omega=1.35$)的迭代计算格式；

(2)判定上述三个迭代格式的收敛性；

(3)取初值 $\boldsymbol{x}^{(0)}=(0,0,0)^{\mathrm{T}}$，用收敛的迭代格式分别求方程组的解，要求

$$\| \boldsymbol{x}^{(k+1)}-\boldsymbol{x}^{(k)} \|_{\infty}<\frac{1}{2}\times 10^{-4}.$$

2. 判断利用 Jacobi 迭代法和 G－S 迭代法求方程组

$$\begin{cases} 20x_1+2x_2+5x_3=24, \\ x_1+10x_2+4x_3=10, \\ 4x_1-3x_2+15x_3=30, \end{cases}$$

的解的收敛性．取初值 $\boldsymbol{x}^{(0)}=(0,0,0)^{\mathrm{T}}$，利用收敛的迭代格式进行计算并精确到小数点后四位．

3. 对于如下线性代数方程组

$$\begin{cases} x_1+2x_2-2x_3=1, \\ x_1+x_2+x_3=2, \\ 2x_1+2x_2+x_3=3, \end{cases}$$

讨论用 Jacobi 迭代和 G－S 迭代的收敛性．

4. 设有方程组 $\boldsymbol{Ax}=\boldsymbol{b}$，其中 $\boldsymbol{A}$ 为对称正定阵，迭代公式为

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{x}^{(k)}+\omega(\boldsymbol{b}-\boldsymbol{Ax}^{(k)}) \quad (k=0,1,2,\cdots).$$

证明当 $0<\omega<\frac{2}{\beta}$ 时上述迭代法收敛(其中 $0<\alpha\leqslant\lambda(\boldsymbol{A})\leqslant\beta$)．

5. 证明矩阵

$$\boldsymbol{A}=\begin{pmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{pmatrix}$$

在 $-\frac{1}{2}<a<1$ 是正定的，而 Jacobi 迭代在 $-\frac{1}{2}<a<\frac{1}{2}$ 时是收敛的．

8 非线性方程求根

8.1 引言

例 8.1 众所周知，理想气体的状态方程为

$$pV = nRT,$$

但是天然气属于真实气体．真实气体的状态参数并不满足理想气体的状态方程，而是存在一定的偏差，偏差的程度通常用偏差因子(Z)来表示，即

$$pV = nZRT$$

式中 p——气体压力，MPa；

V——气体体积，m^3；

T——气体温度，K；

R——气体普适常数，0.008471(MPa·m^3)/(kmol·K)；

Z——气体偏差因子。

1873 年，Vander Waals 从分子热力学理论研究入手，考虑到分子有体积、分子间存在斥力和引力作用这一基本物理现象，根据硬球分子模型，提出了著名的范德华状态方程

$$\left(p+\frac{a}{v^2}\right)(v-b) = RT,$$

其中

$$v = \frac{V}{n}.$$

由于在油气体系气液相平衡计算中需要知道偏差因子，对于 $n=1$mol 分子体系，把经验状态方程 $pV=ZRT$ 代入上式，就得到一个关于偏差因子 Z 的三次方程

$$Z^3-(B+1)Z^2+AZ-AB=0 \tag{8.1}$$

其中

$$A=\frac{ap}{(RT)^2}, B=\frac{bp}{RT}.$$

上述例子表明考虑的因素越多，变量之间的非线性程度越高．

例 8.2 单相管流摩阻系数的计算．

流体在管内流动，往往使部分机械能转化为热能而造成不可逆的能量损失．在单相流动的情况下，不可逆损失主要是摩擦损失．此项损失包括由于流体黏滞性产生的内部损失和管壁形成的外部损失．除层流外，实际的能量损失无法由理论计算确定，而是采用实验的方法和相关分析确定摩阻系数 f．摩阻系数是雷诺数 Re 和相对粗糙度 e/D 的函数．常用的有 Moody 摩阻系数图，摩阻系数 f 与 Re 为双对数关系．

对于摩阻系数的计算，可以利用 Nikurade 的摩阻系数关系式进行计算，即

$$\frac{1}{\sqrt{f}} = 1.74 - 2\ln\frac{2e}{D}.$$

Colebrook 和 White(1939)根据 Moody 图提出了比较完善的关系式

$$\frac{1}{\sqrt{f}} = 1.74 - 2\ln\left(\frac{2e}{D} + \frac{18.7}{Re\sqrt{f}}\right). \tag{8.2}$$

$$Re = \frac{\rho v D}{\mu}$$

式中 e——绝对粗糙度，对于新油管，推荐 e=0.016mm(0.0006in)；

D——管子内径；

Re——雷诺数，它表示流体惯性力与黏滞力之比值，是判别层流与紊流的重要参数．通常认为层流与紊流的分界雷诺数为 2100～2300.

μ——流体黏度．

显然式(8.2)是关于摩阻系数 f 的非线性方程，需用迭代法进行计算．该式可用于紊流的光滑管、过渡区及完全粗糙区．当雷诺数较大时，上式可以转化为 Nikuradse 关系式．

实际上在非线性最小二乘法或最优化问题中，非线性方程及非线性方程组的求解也是经常遇见的．下面讨论在科学和工程计算中经常遇见的求单变量非线性方程

$$f(x) = 0 \tag{8.3}$$

的求根问题．这里 $x\in R, f(x)\in C[a,b]$

方程 $f(x)=0$ 的根 x^* 又称为函数 $f(x)$的零点，它使 $f(x^*)=0$. 若 $f(x)$可以分解为

$$f(x) = (x-x^*)^m g(x),$$

其中 m 为正整数且 $g(x^*)\neq 0$. 当 $m=1$ 时，则称 x^* 为单根；若 $m>1$，则称 x^* 为式(8.3)的 m 重根，或 x^* 为 $f(x)$的 m 重零点．本书中只讨论方程的单根的解法．

当 $f(x)$为代数多项式时，即

$$f(x) = a_0 + a_1 x + \cdots + a_n x^n (a_n \neq 0), \tag{8.4}$$

式(8.3)称为代数方程．根据代数学基本定理，n 次方程 $f(x)=0$ 在复数域内有且仅有 n 个根(包含复根，m 重根算为 m 个根).

当 n=1,2 时，方程的求根问题是简单的．

当 n=3 时，对于给定的方程 $x^3+ax^2+bx+c=0$. 令 $x=y-\frac{a}{3}$，代入得到 $y^3+py+q=0$. 设其根为 y_1, y_2, y_3，则有

$$y_1 = \sqrt[3]{-\frac{q}{2}+\sqrt{\left(\frac{q}{2}\right)^2+\left(\frac{p}{3}\right)^3}} + \sqrt[3]{-\frac{q}{2}-\sqrt{\left(\frac{q}{2}\right)^2+\left(\frac{p}{3}\right)^3}},$$

$$y_2 = \sqrt[3]{-\frac{q}{2}+\sqrt{\left(\frac{q}{2}\right)^2+\left(\frac{p}{3}\right)^3}}\,\omega + \sqrt[3]{-\frac{q}{2}-\sqrt{\left(\frac{q}{2}\right)^2+\left(\frac{p}{3}\right)^3}}\,\omega^2,$$

$$y_3 = \sqrt[3]{-\frac{q}{2}+\sqrt{\left(\frac{q}{2}\right)^2+\left(\frac{p}{3}\right)^3}}\,\omega^2 + \sqrt[3]{-\frac{q}{2}-\sqrt{\left(\frac{q}{2}\right)^2+\left(\frac{p}{3}\right)^3}}\,\omega.$$

其中 $\omega=\dfrac{-1+\sqrt{3}i}{2}$，代入 $x=y-\dfrac{a}{3}$ 从而得到原方程的根 x_1,x_2,x_3.

当 $n=4$ 时，也有一般的求根公式来计算方程的根.

尽管 $n=3,4$ 时有一般的求根公式计算，但是也比较复杂. 而当 $n\geqslant 5$ 时，则没有一般的由方程的系数经有限次四则运算和开方运算求根的方法.

对于一般的代数方程求根都如此复杂，容易想象对于更复杂的超越方程的求根就会更复杂，像式(8.2)就是一例. 下面再看一个简单的超越方程的例子.

例 8.3 考虑如下的非线性方程，其中 α 为参数.

$$f(x)=\mathrm{e}^x-\cos x+\alpha=0$$

方程的根也就是曲线 $y=\mathrm{e}^x+\alpha$ 与曲线 $y=\cos x$ 的交点，通过画图的方法可知：

(1)当 $\alpha\geqslant 1$ 时，两条曲线没有交点，原方程没有根；

(2)当 $-1\leqslant\alpha<1$，两条曲线有无穷多个交点，原方程有无穷多个根；

(3)当 $\alpha<-1$，两条曲线有有限个交点，原方程有有限多个根.

综上所述，要求出一个非线性方程的根的准确值基本上是不可能的，所以转为求方程的根的近似值，这时一般采用迭代法求根. 对于求一个方程的根的近似值，一般需要考虑如下的问题：

(1)有根区间的判定，也就是判定在某个区间内是否有根存在？

(2)如何求出一个满足精度的根？

(3)如何高效地求出一个根？

实际上，上面三个问题是相互联系的.

8.2 有根区间的判定

8.2.1 逐步搜索法

设函数 $f(x)$ 在 $[a,b]$ 上连续，且 $f(a)f(b)<0$，根据连续函数的性质可知方程 $f(x)=0$ 在区间 (a,b) 内一定至少有一个实根，这时称 $[a,b]$ 为方程 $f(x)=0$ 的有根区间. 为明确起见，不妨假定 $f(a)<0,f(b)>0$. 从有根区间 $[a,b]$ 的左端 $x_0=a$ 出发，某个预定的步长 h(例如取 $h=\dfrac{b-a}{N}$，N 为非负整数)一步一步地向右移动，每移动一步进行一次根的“搜索”，即检查节点 $x_k=a+kh$ 上的函数值 $f(x_k)$ 的符号，一旦发现节点 x_k 与节点 x_{k-1} 的函数值异号，即 $f(x_{k-1})f(x_k)<0$，则可以确定一个缩小了的有根区间 $[x_{k-1},x_k]$，其宽度等于预定的步长 h. 这就是逐次搜索法.

例 8.4 考察方程 $f(x)=x^3-11.1x^2+38.8x-41.77=0$ 的有根区间.

解 根据有根区间的定义，对 $f(x)=0$ 的根进行搜索计算，计算结果见表 8.1.

表 8.1 节点函数值符号

x	0	1	2	3	4	5	6
$f(x)$的符号	−	−	+	+	−	−	+

由此可知方程的有根区间为[1,2]、[3,4]和[5,6].

在具体运用上述方法时,步长 h 的选择是个关键. 很明显,只要步长 h 取得足够小,利用这种方法可以得到具有任意精度的近似根. 不过当 h 缩小时,所要搜索的步数相应增多,从而使计算量增大. 因此,如果精度要求比较高,单用这种逐步搜索方法是不合算的. 下述二分法可以看作是逐步搜索方法的一种改进.

8.2.2 二分法

再考察有根区间$[a,b]$,取中点 $x_0=(a+b)/2$ 将它分为两半,然后进行根的搜索,即检查 $f(x_0)$与 $f(a)$是否同号,如果确系同号,说明所求的根 x^* 在 x_0 的右侧,这时令 $a_1=x_0,b_1=b$,否则 x^* 必在 x_0 的左侧,这时令 $a_1=a,b_1=x_0$. 不管出现哪一种情况,新的有根区间$[a_1,b_1]$的长度仅为$[a,b]$的一半.

对压缩了的有根区间$[a_1,b_1]$又可施行同样的手续,即用中点 $x_1=(a_1+b_1)/2$ 将区间$[a_1,b_1]$再分为两半,然后通过根的搜索判定所求的根在 x_1 的哪一侧,从而又确定一个新的有根区间$[a_2,b_2]$,其长度是$[a_1,b_1]$的一半.

如此反复二分下去,即可得出一系列有根区间

$$[a,b]\supset[a_1,b_1]\supset[a_2,b_2]\supset\cdots\supset[a_k,b_k]\supset\cdots$$

其中每个区间都是前一个区间的一半,因此$[a_k,b_k]$的长度 $b_k-a_k=(b-a)/2^k$ 当$k\to\infty$时趋于零,就是说,如果二分过程无限地继续下去,这些区间最终必收缩于一点 x^*,该点显然就是所求的根.

每次二分后,设取有根区间$[a_k,b_k]$的中点 $x_k=(a_k+b_k)/2$ 作为根的近似值,则在二分过程中可以获得一个近似根的序列 $x_0,x_1,x_2,\cdots,x_k,\cdots$该序列必以根 x^* 为极根.

不过在实际计算时,不可能完成这个无限过程,其实也没有这种必要,因为数值分析的结果允许带有一定的误差. 由于

$$|x^*-x_k|\leqslant(b_k-a_k)/2=(b-a)/2^{k+1}$$

只要二分足够多次(即 k 充分大),便有$|x^*-x_k|<\varepsilon$,这里 ε 为预定的精度.

例 8.5 求方程 $f(x)=e^x-1.2-\cos x=0$ 在区间[0,1]内的一个实根.

解 用二分法,取 $a=0,b=1$. 易见函数 $f(x)$连续且 $f(0)=-1.2<0,f(1)>2.7-1-1.2>0$,所以[0,1]是一个有根区间. 取四位有效数字,计算的近似根为 0.6814. 计算结果见表 8.2.

表 8.2 二分法计算结果

a	$f(a)$	b	$f(b)$
0	−1.2000	1.0000	0.9870
0.5000	−0.4289	1.0000	0.9870
0.5000	−0.4289	0.7500	0.1853
0.6250	−0.1427	0.7500	0.1853
0.6250	−0.1427	0.6875	0.0159
0.6563	−0.0647	0.6875	0.0159

续表

a	$f(a)$	b	$f(b)$
0.6719	−0.0248	0.6875	0.0159
0.6797	−0.0045	0.6875	0.0159
0.6797	−0.0045	0.6836	0.0057
0.6797	−0.0045	0.6816	0.0006
0.6807	−0.0020	0.6816	0.0006
0.6812	−0.0007	0.6816	0.0006
0.6814	−0.0001	0.6816	0.0006
0.6814	−0.0001	0.6815	0.0003
0.6814	−0.0001	0.6815	0.0001
0.6814	−0.0001	0.6814	0.0000

二分法对函数 $f(x)$的性质要求不高，只要求连续而已，而且编程简单．它可以得到一个包含近似解的足够小的区间，算法总是收敛的．但是由于它只用到函数值的符号而非函数值本身，所以其收敛速度比较慢．一般用二分法为迭代法提供一个好的近似值而已．

8.3 不动点迭代法

8.3.1 不动点和不动点迭代

为了解方程

$$f(x) = 0 \tag{8.5}$$

类似于线性代数方程组迭代法的构造，把方程式(8.5)改写成等价的形式

$$x = \varphi(x), \tag{8.6}$$

其中 $\varphi(x)$是连续函数．若 x^* 满足 $f(x^*)=0$，则 $x^*=\varphi(x^*)$，反之亦然．

若 $\hat{x}=\varphi(\hat{x})$，则称 $\hat{x}$ 是映射 φ 的不动点．所以求函数 $f(x)$的零点等价于求映射 $\varphi(x)$的不动点．利用式(8.6)自然地构造出不动点迭代公式

$$x_{k+1} = \varphi(x_k) \quad (k = 0,1,2,\cdots). \tag{8.7}$$

其中 x_0 为初始近似值，$\varphi(x)$称为迭代函数．

如果按公式 $x_{k+1}=\varphi(x_k)$确定的数列$\{x_k\}$有极限 $x^*=\lim\limits_{k\to\infty}x_k$，则称迭代过程式(8.7)收敛，且 x^* 为 $\varphi(x)$的不动点，故称式(8.7)为不动点迭代法．

上述迭代法是一种逐次逼近法，其基本思想是将隐式方程式(8.6)归结为一组显式的计算公式(8.7)，就是说，迭代过程实质上是一个逐步显式化的过程．

给定一个非线性方程 $f(x)=0$，可以构造出无穷多种不同的不动点迭代格式．不同的不动点迭代格式虽然在与方程等价这一点是相同的，然而在收敛性以及收敛速度上却可能大相径庭．

例 8.6 求方程 $x^3-6x^2+9x-2=0$ 在区间[3,4]上的一个根.

解 首先将方程转化成如下四种等价形式:

(1)$x=\varphi_1(x)=x^3-6x^2+10x-2$;

(2)$x=\varphi_2(x)=\sqrt{(x^3+9x-2)/6}$;

(3)$x=\varphi_3(x)=x-\dfrac{x^3-6x^2+9x-2}{3x^2-12x+9}$;

(4)$x=\varphi_4(x)=\sqrt[3]{6x^2-9x+2}$.

自然地构造四种相应的不动点迭代格式:

$$x_{k+1}=\varphi_i(x_k)\quad (i=1,2,3,4).$$

对每种迭代格式都取初值 $x_0=3.5$,迭代计算结果见表 8.3.

表 8.3 四种迭代计算结果

k	$x_{k+1}=\varphi_1(x_k)$	$x_{k+1}=\varphi_2(x_k)$	$x_{k+1}=\varphi_3(x_k)$	$x_{k+1}=\varphi_4(x_k)$
0	2.3750	3.4731	3.8000	3.5303
1	1.3027	3.4437	3.7357	3.5571
2	3.0555	3.4115	3.7321	3.5805
3	1.0650	3.3766	3.7321	3.6010
4	3.0526	3.3388		3.6189
5	1.0610	3.2982		3.6345
6	3.0501	3.2548		3.6480
7	1.0577	3.2087		3.6597
8	3.0479	3.1600		3.6698
9	1.0549	3.1090		3.6785
⋮	⋮	⋮		⋮
51	3.0292	⋮		3.7320
⋮	⋮	⋮		
86	1.0257	2.0001		

通过直接计算可知原方程在区间[3,4]内的根是 $2+\sqrt{3}$. 现在分析上述四种迭代法的收敛性问题. 迭代格式(1)产生的计算结果似乎具有振荡性,不收敛;迭代法(2)的计算结果表明近似根越来越偏离[3,4],迭代法不收敛;迭代法(3)迭代了四次就达到了很高的精度,迭代法收敛且收敛很快;迭代法(4)也收敛但迭代次数比迭代法(3)多得多.

迭代法是否收敛和迭代函数有关,下面具体讨论一个迭代法收敛应该满足的条件.

8.3.2 迭代法的收敛性

定理 8.1 假定函数 $\varphi(x)$ 满足下列两个条件:

(1)对任意 $x\in[a,b]$,有

$$a\leqslant\varphi(x)\leqslant b; \tag{8.8}$$

(2)存在正数 $L<1$,使对任意 $x,y\in[a,b]$ 有

$$|\varphi(x)-\varphi(y)|\leqslant L|x-y|,\tag{8.9}$$

则迭代过程 $x_{k+1}=\varphi(x_k)$ 对于任意初值 $x_0\in[a,b]$ 均收敛于方程 $x=\varphi(x)$ 的根 x^*，且有如下的误差估计式：

$$(1)\ |x_k-x^*|\leqslant\frac{1}{1-L}|x_{k+1}-x_k|,\tag{8.10}$$

$$(2)\ |x_k-x^*|\leqslant\frac{L^k}{1-L}|x_1-x_0|.\tag{8.11}$$

证明 若 $\varphi(a)=a$ 或 $\varphi(b)=b$，显然 $\varphi(x)$ 在 $[a,b]$ 存在不动点．下面假设对于 $\forall x\in[a,b]$ $a<\varphi(x)<b$，从而 $\varphi(a)>a$，$\varphi(b)<b$. 如果定义

$$g(x)=\varphi(x)-x,$$

则 $g\in C[a,b]$，且 $g(a)>0$，$g(b)<0$. 由零点定理知 $g(x)$ 在开区间内至少存在一点 x^* 使得 $g(x^*)=0$，即 x^* 是 $\varphi(x)$ 的不动点．如果存在两点 x_1^*、x_2^*，都是 $\varphi(x)$ 的不动点，即 $x_1^*=\varphi(x_1^*)$，$x_2^*=\varphi(x_2^*)$，则由式(8.9)可知

$$|x_1^*-x_2^*|\leqslant L|x_1^*-x_2^*|<|x_1^*-x_2^*|,$$

从而只能得到 $x_1^*=x_2^*$．也就是说映射 $\varphi(x)$ 的不动点 x^* 是唯一的．

注意到

$$|x_k^*-x^*|=|\varphi(x_{k-1})-\varphi(x^*)|\leqslant L|x_{k-1}-x^*|\leqslant\cdots\leqslant L^k|x_0-x^*|,$$

由于 $0<L<1$，故当 $k\to\infty$ 时，序列 $\{x_k\}$ 收敛到 x^*．易见

$$|x_{k+1}-x_k|\leqslant L|x_k-x_{k-1}|.\tag{8.12}$$

对于任意正整数 p，由式(8.12)有

$$\begin{aligned}|x_{k+p}-x_k|&=|x_{k+p}-x_{k+p-1}+x_{k+p-1}-\cdots+x_{k+1}-x_k|\\&\leqslant|x_{k+p}-x_{k+p-1}|+|x_{k+p-1}-x_{k+p-2}|+\cdots+|x_{k+1}-x_k|\\&\leqslant(L^{p-1}+L^{p-2}+\cdots+1)|x_{k+1}-x_k|\\&\leqslant\frac{1}{1-L}|x_{k+1}-x_k|\end{aligned}$$

令 $p\to\infty$，即得式(8.10)．在式(8.10)的基础上反复利用式(8.12)即得式(8.11)．

注：(1)条件式(8.9)称为李普希兹(Lipschitz)条件，L 称为 Lipschitz 常数．由于 L 是小于 1 的正数，所以式(8.9)说明了映射 φ 具有“压缩”性质．实际上定理 8.1 就是泛函分析中经典的“压缩映射定理”在求根中的一个表现形式而已．

(2)判定条件式(8.9)有时比较麻烦，在使用时如果 $\varphi\in C^1[a,b]$ 且对任意 $x\in[a,b]$ 有

$$|\varphi'(x)|\leqslant L<1,\tag{8.13}$$

则由微分中值定理可知对 $\forall x,y\in[a,b]$ 有

$$|\varphi(x)-\varphi(y)|=|\varphi'(\xi)(x-y)|\leqslant L|x-y|,$$

这表明条件式(8.9)可用式(8.13)来代替．

(3)结论式(8.11)是一种先验误差估计，为了达到某种精度，可以用它事先确定迭代步数．而且由式(8.11)也说明正数 L 越小，迭代的次数越少，收敛的越快．

(4)结论式(8.10)表明只要前后两次迭代的误差$|x_{k+1}-x_k|$比较小,则可以说明精确值x^*和近似值x_k之间的误差就比较小. 所以式(8.10)经常用来作为控制迭代终止的条件.

上面给出的对任意$x_0\in[a,b]$,迭代序列的收敛性,通常称为全局收敛性. 事实上很多情况下,全局收敛性不容易检验,所以常常讨论在准确值x^*附近的收敛性问题.

定义 8.1 设$\varphi(x)$在区间$[a,b]$上存在不动点x^*,若存在x^*的某个邻域$N:|x-x^*|\leqslant\delta$,对任意$x_0\in N$,由迭代公式(8.7)产生的迭代序列$\{x_k\}\subset N$且收敛到x^*,则称迭代法式(8.7)局部收敛.

定理 8.2 设x^*为$\varphi(x)$的不动点,$\varphi'(x)$在x^*的邻近连续,且$|\varphi'(x^*)|<1$,则迭代过程式(8.7)是局部收敛的.

证明 由连续函数的性质,存在x^*的某个邻域$N:|x-x^*|\leqslant\delta$,使对于任意$x\in N$成立

$$|\varphi'(x)|\leqslant L<1.$$

此外,对于任意$x\in N$,总有$\varphi(x)\in N$,这是因为

$$|\varphi(x)-x^*|=|\varphi(x)-\varphi(x^*)|\leqslant L|x-x^*|\leqslant|x-x^*|,$$

于是依据定理 8.1 可以断定迭代过程$x_{k+1}=\varphi(x_k)$对于任意初值$x_0\in N$均收敛.

在例 8.6 中,迭代法(3)和(4)产生的迭代序列都收敛,但是迭代的次数有所差异. 为了刻画这种差异,下面讨论迭代序列的收敛速度问题,为此再看下面的一个例子.

例 8.7 用四种迭代格式求方程$x^2-3=0$的正根,方程的准确值是$x^*=\sqrt{3}$.

解 给出四种迭代格式

(1)$x=x^2+x-3$,$\varphi(x)=x^2+x-3$,$x_{k+1}=x_k^2+x_k-3$,这时$\varphi'(x)=2x+1$,$\varphi'(x^*)=2\sqrt{3}+1>1$.

(2)$x=\dfrac{3}{x}$,$\varphi(x)=\dfrac{3}{x}$,$x_{k+1}=\dfrac{3}{x_k}$,这时$\varphi'(x)=-\dfrac{3}{x^2}$,$\varphi'(x^*)=-1$.

(3)$x=x-\dfrac{1}{4}(x^2-3)$,$\varphi(x)=x-\dfrac{1}{4}(x^2-3)$,$x_{k+1}=x_k-\dfrac{1}{4}(x_k^2-3)$,这时$\varphi'(x)=1-\dfrac{x}{2}$,$\varphi'(x^*)=1-\dfrac{\sqrt{3}}{2}\approx0.0134<1$.

(4)$x=\dfrac{1}{2}\left(x+\dfrac{3}{x}\right)$,$\varphi(x)=\dfrac{1}{2}\left(x+\dfrac{3}{x}\right)$,$x_{k+1}=\dfrac{1}{2}\left(x_k+\dfrac{3}{x_k}\right)$,这时$\varphi'(x)=\dfrac{1}{2}\left(1-\dfrac{3}{x^2}\right)$,$\varphi'(x^*)=0$.

在x^*附近取值$x_0=2$,对上述迭代法进行初步的计算可得结果见表 8.4.

表 8.4 迭代计算结果

k	x_k	迭代法 1	迭代法 2	迭代法 3	迭代法 4
0	x_0	2	2	2	2
1	x_1	3	1.5	1.75	1.75
2	x_2	9	2	1.73475	1.732143
3	x_3	87	1.5	1.732361	1.732051
⋮	⋮	⋮	⋮	⋮	⋮

从计算结果可知迭代法(1)和(2)均不收敛,且它们不满足定理 8.2 中的局部收敛性条件.迭代法(3)和(4)均满足定理中的局部收敛性条件,且迭代法(4)比迭代法收敛快些.同理,在例 8.6 中迭代法(3)比迭代法(4)快.为了描述迭代法收敛的快慢,引入收敛阶的概念.

定义 8.2 设迭代过程 $x_{k+1}=\varphi(x_k)$ 收敛于方程 $x=\varphi(x)$ 的根 x^*,误差 $e_k=x_k-x^*$.若存在实数 $p\geqslant 1$ 及非零常数 C 使得

$$\lim_{k\to\infty}\frac{e_{k+1}}{e_k^p}=C$$

则称该迭代过程是 p 阶收敛的.

当 $p=1$ 时,称为线性收敛,$p>1$ 时称为超线性收敛,$p=2$ 时称为平方收敛.

定理 8.3 对于迭代过程 $x_{k+1}=\varphi(x_k)$,如果 $\varphi^{(p)}(x)$ 在所求根 x^* 附近连续且

$$\varphi'(x^*)=\varphi''(x^*)=\cdots=\varphi^{(p-1)}(x^*)=0,\varphi^{(p)}(x^*)\neq 0,$$

则该迭代过程在点 x^* 附近是 p 阶收敛的.

证明 由 $\varphi'(x^*)=0$,从而由定理 8.2 可知迭代法具有局部收敛性.将 $\varphi(x_k)$ 在根 x^* 处做 Taylor 展开,并利用假设条件可知

$$\varphi(x_k)=\varphi(x^*)+\frac{\varphi^{(p)}(\xi)}{p!}(x_k-x^*)^p(\xi\text{在}x_k\text{与}x^*\text{之间}).$$

注意到 $x_{k+1}=\varphi(x_k)$,$x^*=\varphi(x^*)$,由上式得

$$x_{k+1}-x^*=\frac{\varphi^{(p)}(\xi)}{p!}(x_k-x^*)^p.$$

令 $k\to\infty$,有

$$\lim_{k\to\infty}\frac{e_{k+1}}{e_k^p}=C.$$

8.4 牛顿法

8.4.1 牛顿法及其收敛性

牛顿法是求非线性方程根的近似值的重要方法,一般也称作 Newton—Raphson 方法.牛顿法的基本思想就是将非线性方程 $f(x)=0$ 逐步线性化,每一步求线性方程的根.

设函数 $f(x)$ 二阶连续可微,方程 $f(x)=0$ 的根 x^* 的近似值 x_k,将函数 $f(x)$ 在点 x_k 处做 Taylor 展开:

$$f(x)=f(x_k)+f'(x_k)(x-x_k)+\frac{f''(\xi)}{2}(x-x_k)^2,$$

取其前两项,于是方程 $f(x)=0$ 可近似地表示为

$$f(x_k)+f'(x_k)(x-x_k)=0. \tag{8.14}$$

这是一个线性方程.当 $f'(x_k)\neq 0$ 时,记式(8.14)的根为 x_{k+1},则

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \quad (k = 0,1,2,\cdots), \tag{8.15}$$

这就是著名的 Newton 法(或称 Newton－Raphson 法).

牛顿法有明显的几何意义,如图 8.1 所示. 方程 $f(x)=0$ 的根 x^* 可解释为曲线 $y=f(x)$ 与 x 轴的交点的横坐标。

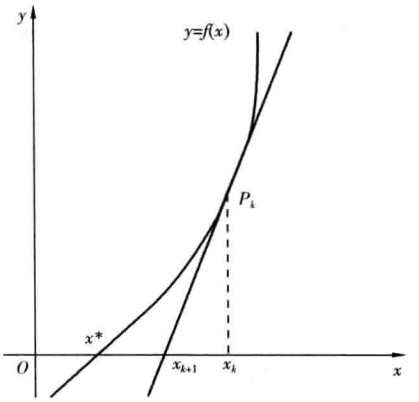

图 8.1　牛顿法的几何意义

设 x_k 是根 x^* 的某个近似值,过曲线 $y=f(x)$ 上横坐标为 x_k 的点 P_k 引切线,并将该切线与 x 轴的交点的横坐标 x_{k+1} 作为 x^* 的新的近似值. 注意到切线方程为

$$y = f(x_k) + f'(x_k)(x - x_k).$$

这样求得的值 x_{k+1} 必满足式(8.14),从而就是牛顿公式(8.15)的计算结果. 由于这种几何背景,牛顿法亦称切线法.

由式(8.15)可知牛顿法的迭代函数为

$$\varphi(x) = x - \frac{f(x)}{f'(x)}.$$

由于

$$\varphi'(x) = \frac{f(x)f''(x)}{[f'(x)]^2},$$

假定 x^* 是 $f(x)=0$ 的一个单根,即 $f(x^*)=0$,则由上式可知 $\varphi'(x^*)=0$,从而由定理 8.3 可知牛顿法在 x^* 附近是平方收敛的.

例 8.8　用牛顿法求方程 $x = e^{-x}$ 的根.

解　将方程转化为 $f(x) = xe^x - 1 = 0$,从而牛顿迭代公式为

$$x_{k+1} = x_k - \frac{x_k - e^{-x_k}}{1 + x_k}.$$

取初值 $x_0=0.5$,计算结果如下

$$x_0 = 0.5, x_1 = 0.57102, x_2 = 0.56716, x_3 = 0.56714, x_4 = 0.56714.$$

若用不动点迭代到同样精度要迭代 17 次之多,可见牛顿迭代法的收敛速度确实是比较快的.

例 8.9 应用牛顿法求$\sqrt{C}$的近似值($C>0$).

解 将求$\sqrt{C}$的近似值转化为求方程 $x^2-C=0$ 的正根. 由牛顿法易见

$$x_{k+1}=\frac{1}{2}\left(x_k+\frac{C}{x_k}\right). \tag{8.16}$$

牛顿法是具有 2 阶精度的局部收敛性方法,但是用牛顿法求 $x^2-C=0$ 的根是一个具有全局收敛的方法. 下面证明这种迭代公式对任意初值 $x_0>0$ 都是收敛的.

由式(8.16)可知

$$x_{k+1}-\sqrt{C}=\frac{1}{2x_k}(x_k-\sqrt{C})^2,$$

$$x_{k+1}+\sqrt{C}=\frac{1}{2x_k}(x_k+\sqrt{C})^2.$$

以上两式相除得

$$\frac{x_{k+1}-\sqrt{C}}{x_{k+1}+\sqrt{C}}=\left(\frac{x_k-\sqrt{C}}{x_k+\sqrt{C}}\right)^2,$$

据此反复递推有

$$\frac{x_k-\sqrt{C}}{x_k+\sqrt{C}}=\left(\frac{x_0-\sqrt{C}}{x_0+\sqrt{C}}\right)^{2^k}, \tag{8.17}$$

记

$$q=\frac{x_0-\sqrt{C}}{x_0+\sqrt{C}},$$

整理式(8.17),得

$$x_k-\sqrt{C}=2\sqrt{C}\,\frac{q^{2^k}}{1-q^{2^k}}.$$

对任意 $x_0>0$,总有$|q|<1$,故由上式推知,当 $k\to\infty$时 $x_k\to\sqrt{C}$,即迭代过程恒收敛.

下面具体地计算$\sqrt{115}$.

取初值 $x_0=10$,利用式(8.17)可得

$$x_0=10, x_1=10.750000, x_2=10.723837, x_3=10.723805, x_4=10.723805.$$

牛顿法的优点是收敛快,缺点是每步迭代要计算 $f(x_k)$和 $f'(x_k)$,这样一来,计算量较大且有时计算 $f'(x_k)$比较困难. 同时对于牛顿法而言,它对初始值 x_0 的选取比较敏感. 初值选取恰当,牛顿法收敛,否则可能发散.

例如,用 Newton 法求解方程 $x^3-x-1=0$. 该方程在 $x=1.5$ 附近的一个根为 x^*. 设取初值 $x_0=1.5$,用牛顿迭代公式为

$$x_{k+1}=x_k-\frac{x_k^3-x_k-1}{3x_k^2-1},$$

计算得到

$$x_1 = 1.34783, x_2 = 1.32520, x_3 = 1.32472.$$

迭代 3 次得到的结果 $x_3=1.32472$ 具有 6 位有效数字.

但是,如果取迭代初值 $x_0=0.6$,则按照牛顿迭代法迭代一次的计算结果为 $x_1=17.9$,这个结果反而比 0.6 更偏离了真实值.

为了克服这些缺点,考虑采用如下一些办法.

8.4.2 简化牛顿法与牛顿下山法

8.4.2.1 简化牛顿法

简化牛顿法就是用初始值 x_0 处的导数来代替 $f'(x_k)$,其迭代公式为

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_0)} \quad (k = 0,1,2,\cdots). \tag{8.18}$$

这种方法计算量少,但是只有线性收敛.

8.4.2.2 牛顿下山法

牛顿法对初值的选取比较敏感,如果初值选取不当,所得 $f(x_k)$ 可能越来越偏离 0. 为了防止发散,在牛顿迭代公式引入参数 λ

$$x_{k+1} = x_k - \lambda \frac{f(x_k)}{f'(x_k)} \quad (k = 0,1,2,\cdots). \tag{8.19}$$

其中 $\lambda(0<\lambda<1)$ 称为下山因子,要求

$$|f(x_{k+1})| < |f(x_k)| \tag{8.20}$$

称为牛顿下山法.

选择下山因子从 $\lambda=1$ 开始,逐次将 λ 折半进行试算,直到能满足下降条件式(8.20)为止.

例 8.10 用牛顿下山法求方程 $f(x)=x^3-x-1=0$ 的根,初值取为 $x_0=0.6$.

解 取初值 $x_0=0.6$,如直接用牛顿法计算,则得到 $x_1=17.9$,偏离了方程的根,且 $|f(17.9)|>|f(x_0)|$. 下面采用牛顿下山法,主要是修正下山因子 λ 的值使下山条件满足. 通过计算可得

取 $\lambda=\frac{1}{2}$,得到

$$x_1^{(1)} = 9.25, |f(9.25)| > |f(x_0)|;$$

取 $\lambda=\frac{1}{2^2}$,得到

$$x_1^{(2)} = 4.925, |f(4.925)| > |f(x_0)|;$$

取 $\lambda=\frac{1}{2^3}$,得到

$$x_1^{(3)} = 2.7625, |f(2.7625)| > |f(x_0)|;$$

取 $\lambda=\frac{1}{2^4}$，得到

$$x_1^{(4)}=1.68125,|f(1.68125)|>|f(x_0)|;$$

取 $\lambda=\frac{1}{2^5}$，得到

$$x_1^{(5)}=1.140625,|f(1.140625)|<|f(x_0)|.$$

这时下山条件已经满足，取 $x_1=1.140625$，以下继续按牛顿法($\lambda=1$)进行迭代

$$x_2=1.140625-\frac{f(1.140625)}{f'(1.140625)}=1.366814,|f(x_2)|<|f(x_1)|,$$

$$x_3=1.366814-\frac{f(1.366814)}{f'(1.366814)}=1.32628,|f(x_3)|<|f(x_2)|,$$

$$x_4=1.32628-\frac{f(1.32628)}{f'(1.32628)}=1.32472,|f(x_4)|<|f(x_3)|,$$

$$x_5=1.32472-\frac{f(1.32472)}{f'(1.32472)}=1.32472.$$

8.5 弦截法

用牛顿法解方程 $f(x)=0$，它在求 x_{k+1} 时不但要求给出函数值 $f(x_k)$，而且要求提供导数值 $f'(x_k)$. 当函数 f 比较复杂时，提供它的导数值往往是有困难的. 现在设法多利用函数值 $f(x_k),f(x_{k-1}),\cdots$ 来回避导数值 $f'(x_k)$ 的计算.

用函数值的差商近似导数得到

$$f'(x_k)\approx\frac{f(x_k)-f(x_{k-1})}{x_k-x_{k-1}},$$

代入牛顿法迭代公式(8.15)得到

$$x_{k+1}=x_k-\frac{(x_k-x_{k-1})f(x_k)}{f(x_k)-f(x_{k-1})}\quad(k=0,1,2,\cdots).\tag{8.21}$$

这就是弦截法(也称为割线法)的迭代公式.

弦截法的几何意义如图 8.2 所示.

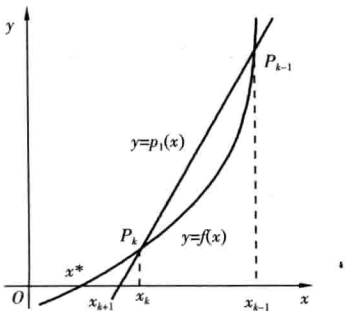

图 8.2 弦截法的几何意义

它不是用 $f(x)$的切线与 x 轴的交点来作为 f 的近似零点，而是通过$(x_{k-1}, f(x_{k-1}))$和$(x_k, f(x_k))$的割线与 x 轴的交点来作为 f 的近似零点．割线的方程为

$$y - f(x_k) = \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}(x - x_k).$$

其零点(曲线与 x 轴交点的横坐标)为

$$x = x_k - \frac{(x_k - x_{k-1}) f(x_k)}{f(x_k) - f(x_{k-1})},$$

把这个零点令为 x_{k+1}，就是弦截法．

弦截法在计算 x_{k+1}时，要用到前两步的近似值 x_{k-1}，x_k 和函数值 $f(x_{k-1})$、$f(x_k)$，因此弦截法必须给出两个初值 x_0 和 x_1．可以证明弦截法是超线性收敛的，收敛阶是

$$p = \frac{1+\sqrt{5}}{2} \approx 1.618.$$

例 8.11 用弦截法解方程 $x^3 - 6x^2 + 9x - 2 = 0$ 在区间$[3,4]$上的近似根．

解 在例 8.6 中实际上已经对该问题进行了讨论，其中的迭代方法(3)实际上就是牛顿迭代法，可以看出牛顿迭代法的收敛速度很快．

取初值 $x_0 = 3$，$x_1 = 4$，利用弦截法的迭代公式(8.21)，计算结果见表 8.5.

表 8.5 弦截法计算结果

k	x_k	有效位
0	3	0
1	4	1
2	3.5	1
3	3.6800	1
4	3.7450	2
5	3.7351	3
6	3.7320	3
7	3.7321	4

同前面的例 8.6 中的牛顿迭代法相比，迭代稍慢，比其他的迭代法快．

8.6 非线性方程组求解

考虑方程组

$$\begin{cases} f_1(x_1, x_2, \cdots, x_n) = 0, \\ f_2(x_1, x_2, \cdots, x_n) = 0, \\ \cdots\cdots \\ f_n(x_1, x_2, \cdots, x_n) = 0. \end{cases} \tag{8.22}$$

其中 $f_1,f_2,\cdots,f_n$ 均为$(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$ 的多元函数．若采用向量记号

$$\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}},\boldsymbol{F}(\boldsymbol{x})=(f_1(x),f_2(x),\cdots,f_n(x))^{\mathrm{T}},$$

则式(8.22)变成了下述向量形式

$$\boldsymbol{F}(\boldsymbol{x})=0. \tag{8.23}$$

当 $n\geqslant 2$,且 $f_i(i=1,2,\cdots,n)$中至少有一个函数是自变量 $x_i(i=1,2,\cdots,n)$的非线性函数时,则称方程组式(8.22)为非线性方程组．

非线性方程组的求解和线性方程组的求解以及非线性方程的求根联系非常紧密,但又有它自己的特点．非线性方程组的求解更加复杂,针对非线性方程组的求解发展了很多方法．下面仅介绍解非线性方程组的牛顿法和两种优化算法即最速下降法和高斯—牛顿法．

8.6.1 牛顿法

牛顿法是求非线性方程组的一种重要方法,可以认为它是前面介绍的非线性方程求根的直接推广．实际上只要把前面介绍的单变量函数 $f(x)$看成向量函数 $\boldsymbol{F}(\boldsymbol{x})$,则可将单变量方程求根方法推广到方程组式(8.23).

若已知方程组式(8.23)的一个近似根 $\boldsymbol{x}^{(k)}=(x_1^{(k)},x_2^{(k)},\cdots,x_n^{(k)})^{\mathrm{T}}$,将函数 $\boldsymbol{F}(\boldsymbol{x})$的分量 $f_i(x)(i=1,2,\cdots,n)$在 $\boldsymbol{x}^{(k)}$处进行多元函数泰勒展开,并取其线性部分,则可以表示为

$$\boldsymbol{F}(\boldsymbol{x})\approx\boldsymbol{F}(\boldsymbol{x}^{(k)})+\boldsymbol{F}'(\boldsymbol{x}^{(k)})(\boldsymbol{x}-\boldsymbol{x}^{(k)}).$$

令上式右端为零,得到线性代数方程组

$$\boldsymbol{F}'(\boldsymbol{x}^{(k)})(\boldsymbol{x}-\boldsymbol{x}^{(k)})=-\boldsymbol{F}(\boldsymbol{x}^{(k)}), \tag{8.24}$$

其中

$$\boldsymbol{F}'(\boldsymbol{x})=\begin{bmatrix}\dfrac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \dfrac{\partial f_1(\boldsymbol{x})}{\partial x_2} & \cdots & \dfrac{\partial f_1(\boldsymbol{x})}{\partial x_n}\\ \dfrac{\partial f_2(\boldsymbol{x})}{\partial x_1} & \dfrac{\partial f_2(\boldsymbol{x})}{\partial x_2} & \cdots & \dfrac{\partial f_2(\boldsymbol{x})}{\partial x_n}\\ \vdots & \vdots & & \vdots\\ \dfrac{\partial f_n(\boldsymbol{x})}{\partial x_1} & \dfrac{\partial f_n(\boldsymbol{x})}{\partial x_2} & \cdots & \dfrac{\partial f_n(\boldsymbol{x})}{\partial x_n}\end{bmatrix}, \tag{8.25}$$

称为 $\boldsymbol{F}(\boldsymbol{x})$的雅可比矩阵．求解线性代数方程组式(8.24),并记解为 $\boldsymbol{x}^{(k+1)}$,则得

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{x}^{(k)}-[\boldsymbol{F}'(\boldsymbol{x}^{(k)})]^{-1}\boldsymbol{F}(\boldsymbol{x}^{(k)})\quad(k=0,1,2,\cdots), \tag{8.26}$$

这就是解非线性方程组式(8.23)的牛顿迭代法．

例 8.12 用牛顿法求解方程组

$$\begin{cases}f_1(x_1,x_2)=x_1+2x_2-3=0,\\ f_2(x_1,x_2)=2x^2+x_2^2-5=0.\end{cases}$$

给定初值 $\boldsymbol{x}^{(0)}=(1.5,1.0)^{\mathrm{T}}$.

解 先求雅可比矩阵

$$\boldsymbol{F}'(\boldsymbol{x})=\begin{pmatrix}1 & 2\\ 4x_1 & 2x_2\end{pmatrix},[\boldsymbol{F}'(\boldsymbol{x})]^{-1}=\frac{1}{2x_2-8x_1}\begin{pmatrix}2x_2 & -2\\ -4x_1 & 1\end{pmatrix},$$

由牛顿法式(8.26)得到

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{x}^{(k)}-\frac{1}{2x_2^{(k)}-8x_1^{(k)}}\begin{pmatrix}2x_2^{(k)} & -2\\ -4x_1^{(k)} & 1\end{pmatrix}\begin{pmatrix}x_1^{(k)}+2x_2^{(k)}-3\\ 2(x_1^{(k)})^2+(x_2^{(k)})^2-5\end{pmatrix},$$

写成分量形式即

$$x_1^{(k+1)}=x_1^{(k)}-\frac{(x_2^{(k)})^2-2(x_1^{(k)})^2+x_1^{(k)}x_2^{(k)}-3x_2^{(k)}+5}{x_2^{(k)}-4x_1^{(k)}},$$

$$x_2^{(k+1)}=x_2^{(k)}-\frac{(x_2^{(k)})^2-2(x_1^{(k)})^2-8x_1^{(k)}x_2^{(k)}+12x_2^{(k)}-5}{2(x_2^{(k)}-4x_1^{(k)})}\quad(k=0,1,2,\cdots).$$

由 $x^{(0)}=(1.5,1.0)^{\mathrm{T}}$ 逐次迭代得到

$$x^{(1)}=(1.5,0.75)^{\mathrm{T}},x^{(2)}=(1.488095,0.755952)^{\mathrm{T}},x^{(3)}=(1.488034,0.755983)^{\mathrm{T}}.$$

$x^{(3)}$ 的每一位都是有效数字.

例 8.12 是一个比较简单的非线性方程组的求解问题,可以求出它的准确解. 这里只是让大家熟悉一下牛顿迭代法的过程.

注意到在牛顿迭代法式(8.26)的过程中,每迭代一步都需要计算一个矩阵 $\boldsymbol{F}'(\boldsymbol{x}^{(k)})$ 的逆矩阵,这种计算量是非常大的,并且如果在某一步矩阵 $\boldsymbol{F}'(\boldsymbol{x}^{(k)})$ 不可逆,则迭代计算就进行不下去了. 另外牛顿迭代法还是存在对初值的敏感性问题. 针对牛顿迭代法的这些问题,发展了许多牛顿迭代法的改进方法如同伦算法、拟牛顿法等.

8.6.2 最速下降法

首先将求非线性方程组的解的问题转化为一个无约束优化问题.

对于非线性方程组式(8.22),构造目标函数

$$\varphi(x)=\frac{1}{2}(\boldsymbol{F}(\boldsymbol{x}))^{\mathrm{T}}\boldsymbol{F}(\boldsymbol{x})=\frac{1}{2}\sum_{i=1}^{n}f_i^2(\boldsymbol{x}).\tag{8.27}$$

由于 $\varphi(x)\geqslant 0$,因此

$$\text{求 }\boldsymbol{x}^*\text{ 使 }\boldsymbol{F}(\boldsymbol{x}^*)=0\Leftrightarrow\text{求 }\boldsymbol{x}^*\text{ 使 }\varphi(\boldsymbol{x}^*)=\min_{x\in R^n}\varphi(\boldsymbol{x}).\tag{8.28}$$

最速下降法是一种搜索方法,它是从目标函数出发,构造使目标函数逐次下降的方法.

搜索方法的一般格式是,对目标函数 $\varphi(\boldsymbol{x}),\boldsymbol{x}\in\boldsymbol{R}^n$,从一个初始值 $\boldsymbol{x}^{(0)}\in\boldsymbol{R}^n$ 出发,由迭代公式

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{x}^{(k)}+\alpha_k\boldsymbol{P}^{(k)}\tag{8.29}$$

产生迭代序列 $\boldsymbol{x}^{(0)},\boldsymbol{x}^{(1)},\cdots,\boldsymbol{x}^{(k)},\cdots$,要求使目标函数值逐步下降,满足

$$|\varphi(\boldsymbol{x}^{(k+1)})-\varphi(\boldsymbol{x}^{(k)})|\leqslant\varepsilon,$$

其中 $\boldsymbol{P}^{(k)} \in \boldsymbol{R}^n$ 是搜索方向，$\alpha_k \in \boldsymbol{R}$ 是搜索步长．

选取搜索方向 $\boldsymbol{P}^{(k)}$，自然是希望使 $\varphi(\boldsymbol{x})$的值下降得越快越好．选取 $\boldsymbol{P}^{(k)}$ 和 α_k 的不同方法就构成了不同的搜索方法．其中最基本、计算最简便的一种搜索方向就是取目标函数的负梯度方向，称为最速下降法．实际上最速下降法是局部下降最快，并不是全局下降最快．因此，最速下降法虽然计算简便但收敛速度较慢．

取 $\varphi(x)$的负梯度方向为搜索方向，即

$$\boldsymbol{P}^{(k)} = -\nabla \varphi(\boldsymbol{x}^{(k)}). \tag{8.30}$$

注意到

$$\nabla\varphi(\boldsymbol{x}^{(k)}) = \left(\frac{\partial\varphi(\boldsymbol{x}^{(k)})}{\partial x_1}, \frac{\partial\varphi(\boldsymbol{x}^{(k)})}{\partial x_2}, \cdots, \frac{\partial\varphi(\boldsymbol{x}^{(k)})}{\partial x_n}\right)^{\mathrm{T}},$$

$$\frac{\partial\varphi(\boldsymbol{x}^{(k)})}{\partial x_i} = f_1(\boldsymbol{x}^{(k)}) \frac{\partial f_1(\boldsymbol{x}^{(k)})}{\partial x_i} + f_2(\boldsymbol{x}^{(k)}) \frac{\partial f_2(\boldsymbol{x}^{(k)})}{\partial x_i} + \cdots + f_n(\boldsymbol{x}^{(k)}) \frac{\partial f_n(\boldsymbol{x}^{(k)})}{\partial x_i}$$

$$= \sum_{j=1}^{n} f_j(\boldsymbol{x}^{(k)}) \frac{\partial f_j(\boldsymbol{x}^{(k)})}{\partial x_i} \quad (i = 1,2,\cdots,n).$$

若记

$$\boldsymbol{P}^{(k)} = (p_1^{(k)}, p_2^{(k)}, \cdots, p_n^{(k)})^{\mathrm{T}},$$

则对 $i=1,2,\cdots,n$，

$$p_i^{(k)} = -\frac{\partial\varphi(\boldsymbol{x}^{(k)})}{\partial x_i} = -\sum_{j=1}^{n} f_j(\boldsymbol{x}^{(k)}) \frac{\partial f_i(\boldsymbol{x}^{(k)})}{\partial x_i}. \tag{8.31}$$

下面讨论如何选取搜索步长 α_k．注意到式(8.29)和式(8.30)，有

$$\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)} = -\alpha_k \nabla\varphi^{(k)}.$$

将函数 $\varphi(\boldsymbol{x}^{(k+1)})$在 $\boldsymbol{x}^{(k)}$ 处进行泰勒展开，忽略高阶项后得到

$$\begin{aligned}\varphi(\boldsymbol{x}^{(k+1)}) &= \varphi(\boldsymbol{x}^{(k)}) + (\boldsymbol{x}^{(k+1)} - x^{(k)})^{\mathrm{T}} \nabla\varphi^{(k)} \\ &= \varphi(\boldsymbol{x}^{(k)}) - \alpha_k (\nabla\varphi(\boldsymbol{x}^{(k)}))^{\mathrm{T}} \nabla\varphi^{(k)} \\ &= \varphi(\boldsymbol{x}^{(k)}) - \alpha_k (\boldsymbol{P}^{(k)}, \boldsymbol{P}^{(k)}),\end{aligned}$$

其中$(\boldsymbol{P}^{(k)}, \boldsymbol{P}^{(k)})$表示两个向量的内积．

最优步长 α_k 的选取应使 $\varphi(\boldsymbol{x}^{(k+1)}) \to 0$，从而得到

$$\alpha_k = \frac{\varphi(\boldsymbol{x}^{(k)})}{(\boldsymbol{P}^{(k)}, \boldsymbol{P}^{(k)})} = \frac{\sum_{i=1}^{n} f_i^2(\boldsymbol{x}^{(k)})}{2\sum_{i=1}^{n} p_i^2(\boldsymbol{x}^{(k)})},$$

从而得到最速下降法的迭代公式为

给定初值 $\boldsymbol{x}^{(0)} \in \boldsymbol{R}^n$，

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \frac{\varphi(\boldsymbol{x}^{(k)})}{(\boldsymbol{P}^{(k)}, \boldsymbol{P}^{(k)})} \boldsymbol{P}^{(k)}, \tag{8.32}$$

其中

$$\boldsymbol{P}^{(k)} = -\nabla\varphi(\boldsymbol{x}^{(k)}),$$

$$(\boldsymbol{P}^{(k)},\boldsymbol{P}^{(k)}) = \sum_{i=1}^{n} p_i^2(\boldsymbol{x}^{(k)}).$$

写成分量形式为

$$x_i^{(k+1)} = x_i^{(k)} + \frac{\sum_{i=1}^{n} f_i^2(\boldsymbol{x}^{(k)})}{2\sum_{i=1}^{n} p_i^2(\boldsymbol{x}^{(k)})} p_i^{(k)} \quad (i = 1,2,\cdots,n;k = 1,2,\cdots). \tag{8.33}$$

例 8.13 用最速下降法解如下非线性方程组，取初值 $\boldsymbol{x}^{(0)} = (1.5,1,1)^{\mathrm{T}}$，迭代一次.

$$\begin{cases} x_1^2 + x_2^2 = 4, \\ x_1^2 - x_2^2 = 1. \end{cases}$$

解 $f_1(\boldsymbol{x}) = x_1^2 + x_2^2 - 4$，$f_2(\boldsymbol{x}) = x_1^2 - x_2^2 - 1$，由式(8.30)可知

$$p_1(\boldsymbol{x}) = -(x_1^2 + x_2^2 - 4)2x_1 - (x_1^2 - x_2^2 - 1)2x_1 = -2x_1(2x_1^2 - 5),$$

$$p_2(\boldsymbol{x}) = -(x_1^2 + x_2^2 - 4)2x_2 - (x_1^2 - x_2^2 - 1)2x_2 = -2x_2(2x_2^2 - 3).$$

采用分量形式(8.33)，对 $k=0$ 有

$$x_1^{(1)} = x_1^{(0)} - \frac{((x_1^{(0)})^2 + (x_2^{(0)})^2 - 4)^2 + ((x_1^{(0)})^2 - (x_2^{(0)})^2 - 1)^2}{2[(2x_1^{(0)}(2x_1^{(0)}x_1^{(0)} - 5))^2 + (2x_2^{(0)}(2x_2^{(0)}x_2^{(0)} - 3))^2]} \times (2x_1^{(0)}(2x_1^{(0)}x_1^{(0)} - 5))$$

$$= 1.5567,$$

$$x_2^{(1)} = x_2^{(0)} - \frac{((x_1^{(0)})^2 + (x_2^{(0)})^2 - 4)^2 + ((x_1^{(0)})^2 - (x_2^{(0)})^2 - 1)^2}{2[(2x_1^{(0)}(2x_1^{(0)}x_1^{(0)} - 5))^2 + (2x_2^{(0)}(2x_2^{(0)}x_2^{(0)} - 3))^2]} \times (2x_2^{(0)}(2x_2^{(0)}x_2^{(0)} - 3))$$

$$= 1.1482.$$

求出解 $\boldsymbol{x}^{(1)} = (1.5567,1,1482)^{\mathrm{T}}$，继续迭代下去可求得 $\boldsymbol{x}^{(2)},\boldsymbol{x}^{(3)},\cdots$. 最终可得到近似解为 $\boldsymbol{x} = (1.5767,1.1782)^{\mathrm{T}}$.

8.6.3 高斯—牛顿法

对于目标函数 $\varphi(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{F}(\boldsymbol{x}))^{\mathrm{T}}\boldsymbol{F}(\boldsymbol{x})$，根据极值存在的必要条件，若 $\boldsymbol{x}$ 是 $\min\limits_{x\in R^n}\varphi(\boldsymbol{x})$ 的解，则 $\boldsymbol{x}$ 必然满足方程组

$$\nabla\varphi(\boldsymbol{x}) = 0, \tag{8.34}$$

其中

$$\nabla\varphi(\boldsymbol{x}) = \sum_{i=1}^{n} f_i(\boldsymbol{x})\,\nabla f_i(\boldsymbol{x})$$

$$= \begin{pmatrix} \dfrac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \dfrac{\partial f_2(\boldsymbol{x})}{\partial x_1} & \cdots & \dfrac{\partial f_n(\boldsymbol{x})}{\partial x_1} \\ \dfrac{\partial f_1(\boldsymbol{x})}{\partial x_2} & \dfrac{\partial f_2(\boldsymbol{x})}{\partial x_2} & \cdots & \dfrac{\partial f_n(\boldsymbol{x})}{\partial x_2} \\ \vdots & \vdots & & \vdots \\ \dfrac{\partial f_1(\boldsymbol{x})}{\partial x_n} & \dfrac{\partial f_2(\boldsymbol{x})}{\partial x_n} & \cdots & \dfrac{\partial f_n(\boldsymbol{x})}{\partial x_n} \end{pmatrix} \begin{pmatrix} f_1(\boldsymbol{x}) \\ f_2(\boldsymbol{x}) \\ \vdots \\ f_n(\boldsymbol{x}) \end{pmatrix} = \boldsymbol{F}'(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{F}(\boldsymbol{x}). \tag{8.35}$$

于是极小化问题就转化为

$$\nabla\varphi(\boldsymbol{x}) = \boldsymbol{F}'(\boldsymbol{x})^{\mathrm{T}} + \boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{0}. \tag{8.36}$$

这是一个不同于原非线性方程组 $\boldsymbol{F}(\boldsymbol{x})=\boldsymbol{0}$ 的新的非线性方程组，一般而言，它比原方程组要复杂得多，求解也更困难.

自然可以考虑采用牛顿迭代法来求解式(8.34)，不过这时需要计算一个 Hessian 矩阵($\boldsymbol{F}(\boldsymbol{x})$的二阶导数)，计算量大. 为了构造更简便的计算方法，先将 $\boldsymbol{F}(\boldsymbol{x})$线性化.

将 $\boldsymbol{F}(\boldsymbol{x})$在点 $\boldsymbol{x}^{(k)}$展开后略去高阶项，有

$$\boldsymbol{F}(\boldsymbol{x}) \approx \boldsymbol{F}(\boldsymbol{x}^{(k)}) + \boldsymbol{F}'(\boldsymbol{x}^{(k)})(\boldsymbol{x}-\boldsymbol{x}^{(k)}) = \boldsymbol{F}'(\boldsymbol{x}^{(k)})\boldsymbol{x} + \boldsymbol{F}(\boldsymbol{x}^{(k)}) - \boldsymbol{F}'(\boldsymbol{x}^{(k)})\boldsymbol{x}^{(k)}.$$

记

$$\boldsymbol{A}_k = \boldsymbol{F}'(\boldsymbol{x}^{(k)}), \boldsymbol{b}_k = \boldsymbol{F}(\boldsymbol{x}^{(k)}) - \boldsymbol{F}'(\boldsymbol{x}^{(k)})\boldsymbol{x}^{(k)},$$

用线性函数 $\boldsymbol{L}_k(\boldsymbol{x})=\boldsymbol{A}_k\boldsymbol{x}+\boldsymbol{b}_k$ 代替非线性函数 $\boldsymbol{F}(\boldsymbol{x})$代入到$\nabla\varphi(\boldsymbol{x})$，注意到式(8.35)得到

$$\nabla\varphi(\boldsymbol{x}) = \boldsymbol{F}'(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{L}'_k(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{L}_k(\boldsymbol{x}) = \boldsymbol{A}_k^{\mathrm{T}}(\boldsymbol{A}_k\boldsymbol{x}+\boldsymbol{b}_k). \tag{8.37}$$

令$\nabla\varphi(\boldsymbol{x}^{(k+1)})=0$，并代入 $\boldsymbol{A}_k$、$\boldsymbol{b}_k$ 得到

$$\boldsymbol{F}'(\boldsymbol{x}^{(k)})^{\mathrm{T}}[\boldsymbol{F}'(\boldsymbol{x}^{(k)})\boldsymbol{x}^{(k+1)} + \boldsymbol{F}(\boldsymbol{x}^{(k)}) - \boldsymbol{F}'(\boldsymbol{x}^{(k)})\boldsymbol{x}^{(k)}] = 0,$$

从而得到

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - [\boldsymbol{F}'(\boldsymbol{x}^{(k)})^{\mathrm{T}}\boldsymbol{F}'(\boldsymbol{x}^{(k)})]^{-1}[\boldsymbol{F}'(\boldsymbol{x}^{(k)})^{\mathrm{T}}\boldsymbol{F}(\boldsymbol{x}^{(k)})], \tag{8.38}$$

这就是高斯—牛顿法. 改变一个写法可以表示为

$$\begin{cases} \boldsymbol{F}'(\boldsymbol{x}^{(k)})^{\mathrm{T}}\boldsymbol{F}'(\boldsymbol{x}^{(k)})\Delta\boldsymbol{x}^{(k)} = -\boldsymbol{F}'(\boldsymbol{x}^{(k)})^{\mathrm{T}}\boldsymbol{F}(\boldsymbol{x}^{(k)}), \\ \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \Delta\boldsymbol{x}^{(k)} \quad (k=0,1,2,\cdots). \end{cases} \tag{8.39}$$

若记

$$\boldsymbol{G}(\boldsymbol{x}^{(k)}) = \boldsymbol{F}'(\boldsymbol{x}^{(k)})^{\mathrm{T}}\boldsymbol{F}'(\boldsymbol{x}^{(k)}),$$

注意到$\nabla\varphi(\boldsymbol{x}^{(k)})=\boldsymbol{F}'(\boldsymbol{x}^{(k)})^{\mathrm{T}}\boldsymbol{F}(\boldsymbol{x}^{(k)})$，从而式(8.39)又可改写成

$$\begin{cases} \boldsymbol{G}(\boldsymbol{x}^{(k)})\Delta\boldsymbol{x}^{(k)} = -\nabla\varphi(\boldsymbol{x}^{(k)}), \\ \boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \Delta\boldsymbol{x}^{(k)} \quad (k=0,1,2,\cdots). \end{cases} \tag{8.40}$$

高斯—牛顿法的计算式不用计算 $\boldsymbol{F}(\boldsymbol{x})$ 的 Hessian 矩阵，只计算 $\boldsymbol{F}'(\boldsymbol{x}^{(k)})$，计算比较简单，而且有较好的收敛性，是一个较好的方法，更常用于求解非线性最小二乘法．

习　题

1. 用二分法求方程 $x^2-x-1=0$ 的正根，并指出其误差．

2. 为求方程 $x^3-x^2-1=0$ 在 1.5 附近的一个根，将方程改写成下列三种等价形式：

(1) $x=1+\dfrac{1}{x^2}$；

(2) $x^3=1+x^2$；

(3) $x^2=\dfrac{1}{x-1}$.

试建立相应的迭代公式并讨论各自的收敛性．选取其中一个公式求出具有 4 位有效数字的近似根．

3. 用下列方法求方程 $f(x)=x^3-3x-1=0$ 在 $x_0=2$ 附近的根．根的准确值 $x^*=1.8793854\cdots$，要求计算结果准确到 4 位有效数字．

(1) 牛顿法 ($x_0=2$)；

(2) 弦截法 ($x_0=2, x_1=1.9$).

4. 用牛顿法求 $f(x)=x-\cos x=0$ 在 $x_0=1$ 附近的实数根，要求满足精度

$$|x_{k+1}-x_k|<0.001.$$

5. 应用牛顿法于方程 $x^3-a=0$，导出求立方根 $\sqrt[3]{a}$ 的迭代公式并讨论其收敛性．

9 常微分方程数值解法

9.1 引言

例 9.1 许多气井都不同程度地含有液体．对于存在底水或边水的气藏，在开采过程中液气比将逐渐增高，会明显地影响气井的产能，甚至将气井淹死．因此，正确地预测气井在较高含液程度下的举升能力，对于气井动态分析和排水采气（如气举）设计具有重要的实际意义．

尽管流体力学的基本方程也适用于油气水多相流动，不过在解决采油或采气工程技术问题时，一般把油水两种液体视为液相，着重考虑气液两相间的作用．描述两相管流的数学模型比单相管流复杂得多．

由于流体的非均质性，在气液两相管流中，气液各相的分布状况可能是多种多样的，存在着各种不同的流动形态，而气液界面又很复杂和多变．因此，寻求实用的、严格的数学解是很困难的．对于采气工程中的气液两相管流，其核心问题是探讨沿程压力损失及影响因素．20世纪六七十年代，一般的处理方法是从物理概念和基本方程出发，采用实验和因次分析法得到描述某一特定两相管流过程的一些无量纲参数，然后根据实验数据得出经验关系式．

Mukherjee 和 Brill(1985)在前人研究工作的基础上，改进实验条件，提出了更为实用的倾斜管（包括水平管）两相流的流型判别准则和应用方便的持液率及摩阻系数经验公式．M－B模型的压降梯度方程为

$$\frac{\mathrm{d}p}{\mathrm{d}z}=\frac{\rho_{\mathrm{m}}g\sin\theta+f_{\mathrm{m}}\rho_{\mathrm{m}}v_{\mathrm{m}}^{2}/2D}{1-\rho_{\mathrm{m}}v_{\mathrm{m}}v_{\mathrm{sg}}/p}$$

$$\rho_{\mathrm{m}}=\rho_{\mathrm{l}}H_{\mathrm{l}}+\rho_{\mathrm{g}}(1-H_{\mathrm{l}})$$

式中 D——油管内径，对于油套环空流动，D 为水力当量直径（套管内径和油管外径之差）；

f_{m}——两相摩阻系数；

ρ_{m}——气液混合物平均密度；

H_{l}——持液率；

v_{sg}——气相表观速度．

若已知起始点 z_0（井口或井底）处的流压 p_0，联合上述方程，就构成了一个常微分方程的初值问题．其一般形式为

$$\begin{cases}\dfrac{\mathrm{d}p}{\mathrm{d}z}=F(z,p),\\ p(z_0)=p_0.\end{cases}\tag{9.1}$$

这类方程有很多数值方法进行求解，经常采用下面即将介绍的具有较高精度的显式四阶龙格—库塔法进行计算．

在科学研究及工程技术领域中，常常会遇到大量的常微分方程如式(9.1)的求解问题．除

了一些简单的方程外，要用传统的数学分析方法找出复杂的变系数或非线性问题的解析表达式是困难的，有时甚至是不可能的．同时许多实际问题也只需要获得解在若干个点上的近似值即可．因此，研究和掌握常常微分方程数值解法，即求出解在一系列离散点上的解的近似值的方法，是很有必要的．

常微分方程数值解法包括常微分方程初值问题数值解法和边值问题数值解法．为介绍数值解法的相关概念和方法的简单起见，本章以常微分方程的两类最简单形式即一阶方程的初值问题和二阶方程的边值问题为例来进行阐述，并着重考察一阶方程的初值问题的数值解法．

所谓一阶常微分方程的初值问题是指

$$\begin{cases} y' = f(x,y), \\ y(x_0) = y_0. \end{cases} \tag{9.2}$$

若函数 $f(x,y)$ 在矩形区域 R：$|x-x_0| \leqslant a$，$|y-y_0| \leqslant b$ 上连续并且关于 y 满足李普希兹(Lipschitz)条件

$$|f(x,y_1) - f(x,y_2)| \leqslant L|y_1 - y_2|, \tag{9.3}$$

其中 $L>0$ 称为李普希兹常数，则初值问题式(9.2)的解 $y=y(x)$ 存在，唯一并且连续依赖于初始条件．

所谓数值解法，即寻找解 $y(x)$ 在一系列离散节点

$$x_0 < x_1 < \cdots < x_n < x_{n+1} < \cdots$$

上的近似值 $y_0, y_1, \cdots, y_n, y_{n+1}, \cdots$（$y_0$ 已知）．相邻两个节点之间的距离 $h_n = x_{n+1} - x_n$ 称为步长．若步长不相等，称将区域进行不等距剖分．为计算简单起见，假定 $h_n = h(n=0,1,2\cdots)$，h 为固定数，这时节点 $x_n = x_0 + nh$，$n=0,1,2,\cdots$．这时称将区域进行等距剖分．

常微分方程初值问题式(9.2)的数值解法的基本特点是它们都采取“步进式”，即求解过程顺着节点排列的次序一步一步向前推进．描述这类算法的基本方法是给出用已知信息 y_n，$y_{n-1}, y_{n-2}, \cdots$ 计算 y_{n+1} 的递推公式．若计算 y_{n+1} 时只用到前一点的值 y_n，这类算法称为单步法．若计算 y_{n+1} 时需要用到 y_{n+1} 前面 k 点的值 $y_n, y_{n-1}, \cdots, y_{n-k+1}$，这类算法称为 k 步法．

9.2 简单的数值方法

9.2.1 欧拉(Euler)法

求解初值问题式(9.2)的数值解的主要手段是寻求对一阶导数的某种离散方法．欧拉法是求解初值问题式(9.2)的一种经典数值方法，它是基于导数的几何意义而建立起来的一种求解式(9.2)一种数值格式．下面从另外的途径来对式(9.2)的导数进行离散．

对式(9.2)中的微分方程从 x_n 到 x_{n+1} 进行积分，得到

$$y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f(t,y(t))\mathrm{d}t. \tag{9.4}$$

右端积分用左矩形公式 $hf(x_n,y(x_n))$ 近似，得到

$$y(x_{n+1}) \approx y(x_n)+hf(x_n,y(x_n)). \tag{9.5}$$

以 y_n 表示 $y(x_n)$ 的近似值，在式(9.5)中用 y_n、y_{n+1} 分别代替 $y(x_n)$ 和 $y(x_{n+1})$，并将“≈”改成“=”，得到

$$y_{n+1} = y_n+hf(x_n,y_n). \tag{9.6}$$

这即是著名的欧拉(Euler)法．由于初值 $y_0=y(x_0)$ 已知，则由式(9.6)可逐步计算出 $y_1=y_0+hf(x_0,y_0)$，$y_2=y_1+hf(x_1,y_1)$，$\cdots$，$y_{n+1}=y_n+hf(x_n,y_n)\cdots$．

若将 $y(x_{n+1})$ 在 x_n 处做 Taylor 展开，有

$$y(x_{n+1}) = y(x_n+h) = y(x_n)+hy'(x_n)+\frac{h^2}{2}y''(\xi_n), \tag{9.7}$$

其中 $\xi_n\in(x_n,x_{n+1})$．注意到 $y'(x_n)=f(x_n,y(x_n))$，略去式(9.7)中的 h^2 项，并用 y_n 近似代替 $y(x_n)$，也可得到式(9.6)．

若将初值问题式(9.2)中的导数 $y'(x_n)$ 用向前差商 $\dfrac{y(x_{n+1})-y(x_n)}{h}$ 代替，这时式(9.2)中的方程可近似写成

$$\frac{y(x_{n+1})-y(x_n)}{h} \approx f(x_n,y(x_n)). \tag{9.8}$$

由此也可以得到 Euler 公式，即式(9.6)．

例 9.2 求解初值问题

$$\begin{cases} y' = y-\dfrac{2x}{y} \quad (0<x<1), \\ y(0) = 1. \end{cases} \tag{9.9}$$

解 初值问题式(9.9)的第一式是一个简单的伯努利(Bernoulli)方程．令 $z=y^2$ 可将式(9.9)转换成关于 z 的线性方程．从而容易得到初值问题的准确解为 $y=\sqrt{2x+1}$．

按照 Euler 公式求数值解，则具体的计算公式为

$$\begin{cases} y_{n+1} = y_n+h\left(y_n-\dfrac{2x_n}{y_n}\right), \\ y_0 = 1. \end{cases} \tag{9.9}$$

取步长 $h=0.1$，这时节点为 $x_n=x_0+nh=nh$ $(n=1,2,\cdots,10)$．计算结果见表 9.1，其中误差 $\varepsilon_n=y(x_n)-y_n$．

表 9.1 欧拉法计算结果

x_n	y_n	$y(x_n)$	$\lvert\varepsilon_n\rvert$	x_n	y_n	$y(x_n)$	$\lvert\varepsilon_n\rvert$
0.1	1.1000	1.0954	0.0046	0.6	1.5090	1.4832	0.0258
0.2	1.1918	1.1832	0.0086	0.7	1.5803	1.5492	0.0311
0.3	1.2774	1.2649	0.0125	0.8	1.6498	1.6125	0.0373
0.4	1.3582	1.3416	0.0166	0.9	1.7178	1.6733	0.0445
0.5	1.4351	1.14142	0.0209	1.0	1.7848	1.7321	0.0527

9.2.2 后退欧拉法

若将式(9.4)中用右矩形公式 $hf(x_{n+1},y_{n+1})$ 近似右端积分，或者将 $y(x_n)$ 在 x_{n+1} 处做 Taylor 展开，或者将式(9.2)中的导数 $y'(x_{n+1})$ 用向后差商 $\frac{y(x_{n+1})-y(x_n)}{h}$，类似地得到求解初值问题式(9.2)的另一种数值计算格式

$$y_{n+1}=y_n+hf(x_{n+1},y_{n+1}) \tag{9.10}$$

称为后退欧拉法.

虽然得到欧拉公式(9.6)和后退欧拉公式(9.10)的方法都一样，但两个公式有着本质的区别. 前者是关于 y_{n+1} 的一个直接计算公式，这种公式称为显式的，有时也称式(9.6)为显式 Euler 公式. 而后者式(9.10)的右端含有未知的 y_{n+1}，除非 $f(x,y)$ 关于 y 线性，一般情况下，式(9.10)是关于 y_{n+1} 的一个非线性方程，这类公式称为隐式的，有时也把式(9.10)称为隐式 Euler 公式.

显式和隐式两类方法各有特点. 考虑到数值计算的稳定性及步长等因素，人们常使用隐式数值格式进行计算，但显式方法一般比隐式方法计算量小得多.

隐式格式式(9.10)通常采用迭代法进行计算，而迭代过程的实质就是将隐式格式逐步显式化.

设用 Euler 公式

$$y_{n+1}^{(0)}=y_n+hf(x_n,y_n)$$

提供迭代初值 $y_{n+1}^{(0)}$，代入式(9.10)右端，使之转化为显式. 直接计算可得

$$y_{n+1}^{(1)}=y_n+hf(x_{n+1},y_{n+1}^{(0)}).$$

然后又用 $y_{n+1}^{(1)}$ 代入式(9.10)得到

$$y_{n+1}^{(2)}=y_n+hf(x_{n+1},y_{n+1}^{(1)}).$$

如此反复进行，有

$$y_{n+1}^{(k+1)}=y_n+hf(x_{n+1},y_{n+1}^{(k)})\quad(k=0,1,2,\cdots). \tag{9.11}$$

假设 $f(x,y)$ 对 y 满足 Lipschitz 条件式(9.3)，由式(9.11)减去式(9.10)得到

$$|y_{n+1}^{(k+1)}-y_{n+1}|=h|f(x_{n+1},y_{n+1}^{(k)})-f(x_{n+1},y_{n+1})|\leqslant hL|y_{n+1}^{(k)}-y_{n+1}|.$$

由此可知，$|y_{n+1}^{(k+1)}-y_{n+1}|\leqslant(hL)^{k+1}|y_{n+1}^{(0)}-y_{n+1}|$. 只要 $hL<1$ 就有 $\lim\limits_{k\to\infty}y_{n+1}^{(k+1)}=y_{n+1}$，即迭代公式(9.11)收敛到解 y_{n+1}.

在应用迭代公式(9.11)进行实际计算时，每迭代一次都要重新计算函数 $f(x,y)$ 的值，而迭代又要反复进行若干次，计算量很大，而且难以预测. 为控制计算量，通常只迭代一两次就转入下一步的计算. 下面给出迭代一次的后退 Euler 公式

$$\begin{cases}\bar{y}_{n+1}=y_n+hf(x_n,y_n),\\ y_{n+1}=y_n+hf(x_{n+1},\bar{y}_{n+1}).\end{cases} \tag{9.12}$$

式(9.12)是显式公式,计算量比 Euler 公式的计算量大,而精度并没有提高,所以在实际计算中一般不采用.

9.2.3 梯形方法与改进的 Euler 公式

在式(9.4)右端积分中用梯形求积公式$\frac{h}{2}[f(x_n,y(x_n))+f(x_{n+1},y(x_{n+1}))]$近似,并用$y_n$、$y_{n+1}$分别代替$y(x_n)$和$y(x_{n+1})$,得到

$$y_{n+1}=y_n+\frac{h}{2}[f(x_n,y_n)+f(x_{n+1},y_{n+1})], \tag{9.13}$$

称为梯形方法.

式(9.13)也可以由 Euler 公式(9.6)和后退 Euler 公式(9.10)两式相加得到.

梯形方法是隐式单步法,实际计算中一般采用迭代法求解. 同后退 Euler 法一样,仍用显式 Euler 公式(9.6)提供迭代初值. 梯形方法的迭代公式为

$$\begin{cases} y_{n+1}^{(0)}=y_n+hf(x_n,y_n), \\ y_{n+1}^{(k+1)}=y_n+\frac{h}{2}[f(x_n,y_n)+f(x_{n+1},y_{n+1}^{(k)}] \quad (k=0,1,2,\cdots). \end{cases} \tag{9.14}$$

类似于 9.2.2 中的讨论,当$\frac{hL}{2}<1$时,仍然有$\lim\limits_{k\to\infty}y_{n+1}^{(k+1)}=y_{n+1}$. 即迭代过程式(9.14)是收敛的.

为简化迭代公式(9.14),可选用显式 Euler 公式求出一个初步的近似值$\bar{y}_{n+1}$,称为预测值. 预测值的精度可能很差,再用梯形公式(9.13)校正一次,即利用式(9.14)迭代一次得到y_{n+1},称为校正值. 这样建立起来的预测—校正格式通常成为改进的欧拉公式.

预测
$$\bar{y}_{n+1}=y_n+hf(x_n,y_n) \tag{9.15}$$

校正
$$y_{n+1}=y_n+\frac{h}{2}[f(x_n,y_n)+f(x_{n+1},\bar{y}_{n+1})] \tag{9.16}$$

或者表示为下列形式

$$\begin{cases} y_p=y_n+hf(x_n,y_n), \\ y_c=y_n+hf(x_{n+1},y_p), \\ y_{n+1}=\frac{1}{2}(y_p+y_c). \end{cases}$$

例 9.3 用改进的欧拉法求解初值问题式(9.9).

解 求解该问题的改进欧拉公式为

$$\begin{cases} y_p = y_n + h\left(y_n - \dfrac{2x_n}{y_n}\right), \\ y_c = y_n + h\left(y_p - \dfrac{2x_{n+1}}{y_p}\right), \\ y_{n+1} = \dfrac{1}{2}(y_p + y_c). \end{cases}$$

仍然取步长 $h=0.1$，其中误差 $\varepsilon_n = y(x_n) - y_n$. 计算结果见表 9.2.

表 9.2　改进欧拉法计算结果

x_n	y_n	$y(x_n)$	$\lvert\varepsilon_n\rvert$	x_n	y_n	$y(x_n)$	$\lvert\varepsilon_n\rvert$
0.1	1.0959	1.0954	0.0005	0.6	1.4860	1.4832	0.0028
0.2	1.1841	1.1832	0.0009	0.7	1.5525	1.5492	0.0033
0.3	1.2662	1.2649	0.0013	0.8	1.6153	1.6125	0.0028
0.4	1.3434	1.3416	0.0018	0.9	1.6782	1.6733	0.0049
0.5	1.4164	1.14142	0.0022	1.0	1.7379	1.7321	0.0058

表 9.1 中的 ε_n 的最小值是 4.6×10^{-3}，ε_n 的最大值是 5.27×10^{-2}. 表 9.2 中 ε_n 的最小值是 5×10^{-4}，ε_n 的最大值是 5.8×10^{-3}. 对比表 9.1 和表 9.2 可知，改进的欧拉法的计算精度明显提高.

例 9.4　用 Euler 方法、后退的 Euler 方法、梯形方法和改进的 Euler 法求常微分方程初值问题

$$\begin{cases} y' = -y + x + 1 \quad \left(0 < x < \dfrac{1}{2}\right), \\ y(0) = 1. \end{cases}$$

的解.

解　初值问题中第一式是一个简单的线性非齐次方程，易知初值问题的准确解为

$$y(x) = \mathrm{e}^{-x} + x.$$

取步长 $h=0.1$，这时 $x_n = nh\ (n=1,2,3,4,5)$. 由于 $f(x,y) = -y+x+1$ 对 y 线性，所以后退 Euler 法和梯形方法不用迭代方法. 经过简单的计算可知，求解上述初值问题的四种数值格式分别为：

Euler 方法　　$y_{n+1} = 0.9y_n + 0.1x_n + 0.1$,

后退的 Euler 方法　　$y_{n+1} = \dfrac{y_n + 0.1x_n + 0.11}{1.1}$,

梯形方法　　$y_{n+1} = \dfrac{0.95y_n + 0.1x_n + 0.105}{1.05}$,

改进的 Euler 法　　$y_{n+1} = 0.905y_n + 0.095x_n + 0.1$.

计算结果和相应的误差分别见表 9.3 和表 9.4.

表 9.3　四种迭代法计算结果

x_n	Euler 法 $\lvert y_{n+1}\rvert$	后退 Euler 法 $\lvert y_{n+1}\rvert$	梯形法 $\lvert y_{n+1}\rvert$	改进的 Euler 法 $\lvert y_{n+1}\rvert$	准确值 $\lvert y_{n+1}\rvert$
0.1	1.000000	1.009091	1.004762	1.005000	1.004837
0.2	1.010000	1.026446	1.018594	1.019025	1.018731
0.3	1.029000	1.051315	1.040633	1.041218	1.040818
0.4	1.056100	1.083013	1.070096	1.070802	1.070320
0.5	1.090490	1.120921	1.106278	1.107076	1.106531

表 9.4　计算结果误差

x_n	Euler 法 $\lvert \varepsilon_n\rvert$	后退 Euler 法 $\lvert \varepsilon_n\rvert$	梯形法 $\lvert \varepsilon_n\rvert$	改进的 Euler 法 $\lvert \varepsilon_n\rvert$
0.1	0.004837	0.004254	0.000075	0.000163
0.2	0.008731	0.007715	0.000137	0.000294
0.3	0.011818	0.010497	0.000185	0.000400
0.4	0.014220	0.012693	0.000224	0.000482
0.5	0.016041	0.014390	0.000253	0.000545

从上面的数值计算结果可以看出，梯形法和改进的 Euler 法的误差较小，Euler 法和后退 Euler 法的误差较大．为刻划单步法的误差，引入截断误差、阶以及收敛性等概念．

9.2.4　单步法的有关概念

初值问题式(9.2)的单步法公式的一般形式为

$$y_{n+1} = y_n + h\varphi(x_n, y_n, y_{n+1}, h), \tag{9.17}$$

其中多元函数 φ 与 $f(x,y)$ 有关．当 φ 含有 y_{n+1} 时，方法是隐式的，当 φ 不含有 y_{n+1} 时，方法是显式的．所以一般显式单步法可表示为

$$y_{n+1} = y_n + h\varphi(x_n, y_n, h), \tag{9.18}$$

$\varphi(x,y,h)$ 称为增量函数．例如对 Euler 法式(9.6)，$\varphi(x,y,h)=f(x,y)$.

从例 9.4 可以看出，不同的方法求出的 y_n 与准确解 $y(x_n)$ 的误差是不同的．称

$$e_n = y(x_n) - y_n.$$

为某方法在点 x_n 处的整体截断误差．显然 e_n 不仅与节点 x_n 处的计算有关，而且与前面节点的计算有关．为分析误差方便，引入显式单步法的局部截断误差的概念．

定义 9.1　设 $y(x_n)$ 是初值问题式(9.2)的准确解．称

$$T_{n+1} = y(x_{n+1}) - y(x_n) - h\varphi(x_n, y(x_n), h) \tag{9.19}$$

为显式单步法式(9.18)在节点 x_{n+1} 处的局部截断误差．

T_{n+1} 之所以称为局部的，是因为假设在 x_n 前各步没有误差，当 $y_n=y(x_n)$ 时，计算一步而产生的整体截断误差．易见 $y_n=y(x_n)$ 时

$$\begin{aligned} y(x_{n+1}) - y_{n+1} &= y(x_{n+1}) - [y_n + h\varphi(x_n, y_n, h)] \\ &= y(x_{n+1}) - y(x_n) - h\varphi(x_n, y(x_n), h) \\ &= T_{n+1}. \end{aligned}$$

定义 9.2 设 $y(x)$是初值问题式(9.2)的准确解．若存在最大整数 p，使显式单步法式(9.18)的局部截断误差满足

$$T_{n+1} = y(x_n + h) - y(x_n) - h\varphi(x_n, y(x_n), h) = O(h^{p+1}), \tag{9.20}$$

则称式(9.18)具有 p 阶精度，或称式(9.18)的阶是 p.

若将式(9.20)在 x_n 处做 Taylor 展开，可以写成

$$T_{n+1} = \psi(x_n, y(x_n))h^{p+1} + O(h^{p+2}), \tag{9.21}$$

称 $\psi(x_n, y(x_n))h^{p+1}$ 为局部截断误差主项．

例 9.5 对于 Euler 方法，由 Taylor 展开式有

$$\begin{aligned} T_{n+1} &= y(x_{n+1}) - y(x_n) - hf(x_n, y(x_n)) \\ &= y(x_n + h) - y(x_n) - hy'(x_n) \\ &= \frac{h^2}{2}y''(x_n) + O(h^3), \end{aligned}$$

从而 Euler 方法是 1 阶方法，局部截断误差主项为 $\frac{h^2}{2}y''(x_n)$.

以上关于局部截断误差的定义对于隐式单步法式(9.17)也是适用的．

例 9.6 对于后退 Euler 法，其局部截断误差为

$$\begin{aligned} T_{n+1} &= y(x_{n+1}) - y(x_n) - hf(x_{n+1}, y(x_{n+1})) \\ &= hy'(x_n) + \frac{h^2}{2}y''(x_n) + O(h^3) - hy'(x_{n+1}) \\ &= hy'(x_n) + \frac{h^2}{2}y''(x_n) + O(h^3) - h[y'(x_n) + hy''(x_n) + O(h^2)] \\ &= -\frac{h^2}{2}y''(x_n) + O(h^3), \end{aligned}$$

从而后退 Euler 方法是 1 阶方法，局部截断误差主项是 $-\frac{h^2}{2}y''(x_n)$.

例 9.7 对于梯形方法，其局部截断误差为

$$\begin{aligned} T_{n+1} &= y(x_{n+1}) - y(x_n) - \frac{h}{2}[f(x_n, y(x_n)) + f(x_{n+1}, y(x_{n+1}))] \\ &= y(x_{n+1}) - y(x_n) - \frac{h}{2}[y'(x_n) + y'(x_{n+1})] \\ &= -\frac{h^3}{12}y'''(x_n) + O(h^4), \end{aligned}$$

从而梯形方法是 2 阶方法，其局部截断误差主项是 $-\frac{h^3}{12}y'''(x_n)$.

对于单步法，除了讨论方法的截断误差外，还有必要讨论其收敛性和稳定性.

定义 9.3 对于单步法式(9.18)，它在 $x_n=x_0+nh$ 处的解为 y_n. 若对任意固定的 x_n，均有 $\lim\limits_{h\to 0}y_n=y(x_n)$，则称单步法式(9.18)是收敛的.

类似的定义可用于隐式单步法. 显然若方法是收敛的，在固定点 x_n 处的整体截断误差 $e_n=y(x_n)-y_n\to 0$.

定理 9.1 设单步法式(9.18)具有 $p(\geqslant 1)$ 阶精度，且增量函数 $\varphi(x,y,h)$ 关于 y 满足 Lipschitz 条件，即存在常数 $L_\varphi>0$，使得对任意 y_1,y_2

$$|\varphi(x,y_1,h)-\varphi(x,y_2,h)|\leqslant L_\varphi|y_1-y_2| \tag{9.22}$$

成立. 又假设初值 y_0 是准确的，即 $y_0=y(x_0)$. 则式(9.18)收敛且 $y(x_n)-y_n=O(h^p)$.

证明 注意到 $e_n=y(x_n)-y_n$，由局部截断误差的定义有

$$y(x_{n+1})=y(x_n)+h\varphi(x_n,y(x_n),h)+T_{n+1},$$

将上式和式(9.18)相减得到

$$e_{n+1}=e_n+h[\varphi(x_n,y(x_n),h)-\varphi(x_n,y_n,h)]+T_{n+1}.$$

由于所给方法具有 p 阶精度，从而由定义 9.2 知存在常数 $C>0$，使得

$$|T_{n+1}|\leqslant Ch^{p+1},$$

从而有

$$|e_{n+1}|\leqslant|e_n|+h|[\varphi(x_n,y(x_n),h)-\varphi(x_n,y_n,h)]|+Ch^{p+1}.$$

注意到条件式(9.22)有

$$|e_{n+1}|\leqslant(1+hL_\varphi)|e_n|+Ch^{p+1}, \tag{9.23}$$

由此不等式反复递推可得

$$|e_n|\leqslant(1+hL_\varphi)^n|e_0|+\frac{Ch^p}{L_\varphi}[(1+hL_\varphi)^n-1]. \tag{9.24}$$

注意到当 $x_n-x_0=nh\leqslant T$ 时候，$(1+hL_\varphi)^n\leqslant(e^{hL_\varphi})^n\leqslant e^{TL_\varphi}$，由此得到

$$|e_n|\leqslant|e_0|e^{TL_\varphi}+\frac{Ch^p}{L_\varphi}[e^{TL_\varphi}-1]. \tag{9.25}$$

故当 $y_0=y(x_0)$ 即 $e_0=0$ 时，结论成立.

根据该定理，判定单步法的收敛性，就归结为验证增量函数 φ 能否满足 Lipschitz 条件式(9.22). 不难验证 Euler 法与改进的 Euler 法均是收敛的.

单步法的稳定性是指数值格式的求解过程中的计算误差在传播过程中是否会恶性增长的问题. 在实际计算中，若某一步计算产生的计算误差在后面的计算中能够被控制，甚至是衰减的，则称方法是稳定的. 对于微分方程初值问题，稳定性不但与方法本身有关，也与步长 h 的大小有关，而且也与方程中的 $f(x,y)$ 有关. 为了只考察数值方法本身，通常选用模型方程

$$y'=\lambda y(\lambda<0) \tag{9.26}$$

来检验数值方法的稳定性．这时某种数值方法的稳定性条件实际上就是对步长 h 的限制．

用单步法求解上述模型方程式(9.26)，其数值格式为

$$y_{n+1} = E(h\lambda)y_n. \tag{9.27}$$

例如用 Euler 法和后退 Euler 法求解式(9.26)的数值格式分别为

$$y_{n+1} = (1+h\lambda)y_n$$

和

$$y_{n+1} = \frac{1}{1-h\lambda}y_n,$$

这时 $E(h\lambda)$分别为$(1+h\lambda)$和$\frac{1}{1-h\lambda}$.

定理 9.2 单步法式(9.17)用于解模型方程式(9.26)，若满足 $|E(h\lambda)|<1$，则称式(9.17)是(绝对)稳定的．

9.3 显式龙格—库塔方法

9.3.1 显式龙格—库塔方法的一般形式

显式单步法最简单的方法就是 Euler 法，但其精度只有 1 阶．改进的 Euler 法也是一种显式单步法，其精度是 2 阶．构造高阶的显式单步格式，由式(9.2)得到

$$y(x_{n+1}) - y(x_n) = \int_{x_n}^{x_{n+1}} f(x,y(x))\mathrm{d}x \tag{9.28}$$

要使公式的精度提高，就必须使右端积分的数值求积公式的精度提高．这时必然要增加求积节点，因此可将式(9.28)右端用求积公式表示为

$$\int_{x_n}^{x_{n+1}} f(x,y(x))\mathrm{d}x \approx h\sum_{i=1}^{r} c_i f(x_n+\lambda_i h, y(x_n+\lambda_i h)).$$

一般而言，r 越大，精度越高．为得到便于计算的显式方法，将公式表示为

$$y_{n+1} = y_n + h\varphi(x_n,y_n,h), \tag{9.29}$$

其中

$$\varphi(x_n,y_n,h) = \sum_{i=1}^{r} c_i K_i, \tag{9.30}$$

$$K_1 = f(x_n,y_n),$$

$$K_i = f(x_n+\lambda_i h, y_n + h\sum_{j=1}^{i-1}\mu_{ij}K_i) \quad (i=2,3,\cdots,r).$$

其中 c_i,λ_i,μ_{ij} 均为常数，整数 $r\geqslant 1$．式(9.30)用了 r 个 K_i 值，称式(9.29)和式(9.30)为 r 级显式龙格—库塔(Runge—Kutta)法，简称 R—K 法．

当 $r=1,\varphi(x_n,y_n,h)=f(x_n,y_n)$ 时，就是欧拉法，这时方法的精度为 1 阶．当 $r=2$ 时，可以证明改进的欧拉法是其中的一种，下面将要证明这一点．现在就 $r=2$ 的情况，借助于 Taylor 展开法具体推导 R－K 公式，并给出 $r=3,4$ 时的显式 R－K 公式．

9.3.2 二阶显式 R－K 方法

$r=2$，即 2 阶 R－K 公式的一般形式为

$$\begin{cases} y_{n+1}=y_n+h(c_1K_1+c_2K_2), \\ K_1=f(x_n,y_n), \\ K_2=f(x_n+\lambda_2 h,y_n+\mu_{21}hK_1), \end{cases} \tag{9.31}$$

其中 $c_1,c_2,\lambda_2,\mu_{21}$ 均为待定常数．下面将适当选取系数，使公式的阶数尽可能高．式(9.31)的局部截断误差为

$$T_{n+1}=y(x_{n+1})-y(x_n)-h[c_1f(x_n,y_n)+c_2f(x_n+\lambda_2h,y_n+\mu_{21}hf_n)]. \tag{9.32}$$

这里 $y_n=y(x_n),f_n=f(x_n,y_n)$．将 $y(x_{n+1})$ 在 $x=x_n$ 处做 Taylor 展开，得到

$$y(x_{n+1})=y(x_n+h)=y_n+hy'_n+\frac{h^2}{2}y''_n+\frac{h^3}{3!}y'''_n+O(h^4). \tag{9.33}$$

由 $y'=f(x,y)$ 以及二元函数求偏导数的链式法则可知

$$y'_n=f(x_n,y_n)=f_n,$$

$$y''_n=\frac{\mathrm{d}}{\mathrm{d}x}f(x_n,y(x_n))=f'_x(x_n,y_n)+f'_y(x_n,y_n)f_n,$$

$$\begin{aligned} y'''_n&=\frac{\mathrm{d}}{\mathrm{d}x}[f'_x(x_n,y_n)+f'_y(x_n,y_n)f_n] \\ &=f''_{xx}(x_n,y_n)+2f_nf''_{xy}(x_n,y_n)+f_n^2f''_{yy}(x_n,y_n)+f'_y(x_n,y_n)[f'_x(x_n,y_n)+f'_y(x_n,y_n)f_n]. \end{aligned} \tag{9.34}$$

将式(9.32)中的 f 在 (x_n,y_n) 处做二元 Taylor 展开，得到

$$f(x_n+\lambda_2h,y_n+\mu_{21}hf_n)=f_n+\lambda_2hf'_x(x_n,y_n)+\mu_{21}hf_nf'_y(x_n,y_n)+O(h^2). \tag{9.35}$$

将式(9.33)、式(9.34)和式(9.35)代入式(9.32)，化简得到

$$\begin{aligned} T_{n+1}&=hf_n+\frac{h^2}{2}[f'_x(x_n,y_n)+f'_y(x_n,y_n)f_n] \\ &\quad -h[c_1f_n+c_2(f_n+\lambda_2hf'_x(x_n,y_n)+\mu_{21}hf_nf'_y(x_n,y_n))]+O(h^3) \\ &=(1-c_1-c_2)f_nh+\left(\frac{1}{2}-c_2\lambda_2\right)f'_x(x_n,y_n)h^2+\left(\frac{1}{2}-c_2\mu_{21}\right)f'_y(x_n,y_n)f_nh^2+O(h^3). \end{aligned}$$

要使式(9.31)具有二阶精度，必须使

$$1-c_1-c_2=0,\frac{1}{2}-c_2\lambda_2=0,\frac{1}{2}-c_2\mu_{21}=0. \tag{9.36}$$

显然式(9.36)的解是不唯一的.

若令 $c_2=a\neq 0$,则有

$$c_1=1-a,\lambda_2=\mu_{21}=\frac{1}{2a},$$

这样得到的公式称为二级二阶 R－K 方法.

若取 $a=\frac{1}{2}$,则 $c_1=c_2=\frac{1}{2},\lambda_2=\mu_{21}=1$,即改进的欧拉法.

若取 $a=1$,则 $c_1=0,c_2=1,\lambda_2=\mu_{21}=\frac{1}{2}$. 得到计算公式

$$\begin{cases} y_{n+1}=y_n+hK_2, \\ K_1=f(x_n,y_n), \\ K_2=f\left(x_n+\frac{h}{2},y_n+\frac{h}{2}K_1\right), \end{cases} \tag{9.37}$$

称为中点公式.

或者写成

$$y_{n+1}=y_n+hf\left(x_n+\frac{h}{2},y_n+\frac{h}{2}f(x_n,y_n)\right).$$

由上述计算过程可以推出,$r=2$ 的显式 R－K 方法的阶只能是 2,而不能得到 3 阶公式.

9.3.3 三阶与四阶显式 R－K 方法

对于 $r=3$ 的情形,式(9.29)和式(9.30)可以表示为

$$\begin{cases} y_{n+1}=y_n+h(c_1K_1+c_2K_2+C_3K_3), \\ K_1=f(x_n,y_n), \\ K_2=f(x_n+\lambda_2h,y_n+\mu_{21}hK_1), \\ K_3=f(x_n+\lambda_3h,y_n+\mu_{31}hK_1+\mu_{32}hK_2). \end{cases} \tag{9.38}$$

其中,c_1,c_2,c_3 和 $\lambda_2,\mu_{21},\lambda_3,\mu_{31},\mu_{32}$ 均为待定常数. 利用局部截断误差

$$T_{n+1}=y(x_{n+1})-y(x_n)-h(c_1K_1+c_2K_2+c_3K_3)$$

将 K_2,K_3 在 (x_n,y_n) 处做 Taylor 展开,并使 $T_{n+1}=O(h^4)$,可以得到

$$\begin{cases} c_1+c_2+c_3=1, \\ \lambda_2=\mu_{21}, \\ \lambda_3=\mu_{31}+\mu_{32}, \\ c_2\lambda_2+c_3\lambda_3=\frac{1}{2}, \\ c_2\lambda_2^2+c_3\lambda_3^2=\frac{1}{3}, \\ c_3\lambda_2\mu_{32}=\frac{1}{6}. \end{cases} \tag{9.39}$$

这是 8 个未知数 6 个方程的方程组，解也是不唯一的．常见的一种三级三阶方法是下面的 Kutta 三阶方法，即

$$\begin{cases} y_{n+1} = y_n + \dfrac{h}{6}(K_1 + 4K_2 + K_3), \\ K_1 = f(x_n, y_n), \\ K_2 = f\left(x_n + \dfrac{h}{2}, y_n + \dfrac{h}{2}K_1\right), \\ K_3 = f(x_n + h, y_n - hK_1 + 2hK_2). \end{cases}$$

对于 $r=4$，最常用的是下面的经典 Runge－Kutta 方法，即

$$\begin{cases} y_{n+1} = y_n + \dfrac{h}{6}(K_1 + 2K_2 + 2K_3 + K_4), \\ K_1 = f(x_n, y_n), \\ K_2 = f\left(x_n + \dfrac{h}{2}, y_n + \dfrac{h}{2}K_1\right), \\ K_3 = f\left(x_n + \dfrac{h}{2}, y_n + \dfrac{h}{2}K_2\right), \\ K_4 = f(x_n + h, y_n + hK_3). \end{cases}$$

经典的 4 阶 R－K 方法精度高，但它的计算量也很大．在同样步长的情形下，Euler 方法每步只要计算一个函数值，而经典的 R－K 方法要计算四个函数值．下面的例子中 Euler 方法用步长 h_1，二阶的改进 Euler 法用步长 $2h_1$，而四阶的经典方法用步长 $4h_1$．这样从 x_n 到 $x_n + 4h_1$ 三种方法都计算了四个函数值，计算量大体相当．

例 9.8 用 Euler 方法（$h=0.025$）、改进的 Euler 法（$h=0.05$）和经典的 R－K 方法（$h=0.1$）求初值问题

$$\begin{cases} y' = -y + 1, \\ y(0) = 0, \end{cases}$$

的解．

解 计算结果见表 9.5.

表 9.5 几种迭代法的计算结果

x_n	Euler 方法 $h=0.025$	改进的 Euler 法 $h=0.05$	经典的 R－K 方法 $h=0.1$	准确值 $y(x_n)$
0	0	0	0	0
0.1	0.096312	0.095123	0.09516250	0.09516258
0.2	0.183348	0.181193	0.18126910	0.18126925
0.3	0.262001	0.259085	0.25918158	0.25918178
0.4	0.333079	0.329563	0.32967971	0.32967995
0.5	0.397312	0.393337	0.39346906	0.39346934

从计算结果来看，在工作量大致相同的情况下，经典 R－K 方法比其他两种方法的结果好的多．在 $x=0.5$，三种方法的误差分别是 3.8×10^{-3}、1.3×10^{-4} 和 2.8×10^{-7}．

除了显式 R－K 方法，还有隐式 R－K 方法．

9.4 线性多步法

常微分方程初值问题式(9.2)的数值解法中除了 Euler 法、R－K 方法等单步法外，还有一种类型的数值方法．即某一步解的公式 y_{n+1} 不仅与前一步的解的值 y_n 有关，而且与前若干步解的值 $y_{n-1}, y_{n-2}, \cdots, y_{n-k}$ $(k\leqslant n)$ 都有关，这就是多步法．如果充分利用前面多步的信息来预测 y_{n+1}，则可以期望获得较高的精度．

构造多步法的主要途径是基于数值积分的方法和基于 Taylor 展开法．前者是将常微分方程两端积分后利用插值求积公式得到，后者是利用局部截断误差定义和 Taylor 展开得到．后面将介绍如何利用这两种方法构造多步法公式．

9.4.1 线性多步法的一般公式

一般的多步法公式可表示为

$$y_{n+k}=\sum_{i=0}^{k-1}\alpha_i y_{n+i}+h\sum_{i=0}^{k}\beta_i f_{n+i}. \tag{9.40}$$

其中，y_{n+i} 为 $y(x_{n+i})$ 的近似，$f_{n+i}=f(x_{n+i},y_{n+i})$，$x_{n+i}=x_n+ih$；$\alpha_i,\beta_i$ 为常数；α_0,β_0 不全为零．由于式(9.40)给出了 y_{n+i} 和 $f_{n+i}(i=0,1,2,\cdots,k)$ 之间的线性关系，故称式(9.40)为线性 k 步法．

在利用式(9.40)实际计算时，需用单步法提供 k 个初值 $y_0, y_1\cdots, y_{k-1}$，再由式(9.40)逐次求出 $y_k, y_{k+1}, \cdots$．若 $\beta_k=0$，称式(9.40)为显式 k 步法，这时 y_{n+k} 可直接由式(9.40)计算得到．若 $\beta_k\neq0$，称式(9.40)为隐式 k 步法，这时求解 y_{n+k} 一般需要采用迭代法计算．式(9.40)中系数 α_i、β_i 可根据方法的局部截断误差以及阶确定．

定义 9.4 设 $y(x)$ 是初值问题式(9.2)的准确解，则线性多步法式(9.40)在 x_{n+k} 处的局部截断误差为

$$T_{n+k}=y(x_{n+k})-\sum_{i=0}^{k-1}\alpha_i y(x_{n+i})-h\sum_{i=0}^{k}\beta_i y'(x_{n+i}). \tag{9.41}$$

若 $T_{n+k}=O(h^{p+1})$，则称式(9.40)是 p 阶的．

将 T_{n+k} 在 x_n 处做 Taylor 展开，注意到

$$y(x_{n+i})=y(x_n+ih)=y(x_n)+ihy'(x_n)+\frac{(ih)^2}{2}y''(x_n)+\frac{(ih)^3}{3!}y'''(x_n)+\cdots$$

$$y'(x_{n+i})=y'(x_n+ih)=y'(x_n)+ihy''(x_n)+\frac{(ih)^2}{2}y'''(x_n)+\cdots$$

代入式(9.41)得到

$$T_{n+k}=c_0y(x_n)+c_1hy'(x_n)+c_2h^2y''(x_n)+\cdots+c_ph^py^{(p)}(x_n)+\cdots \tag{9.42}$$

其中

$$\begin{aligned}
c_0 &= 1-(\alpha_0+\alpha_1+\cdots+\alpha_{k-1}),\\
c_1 &= k-[\alpha_1+2\alpha_2+\cdots+(k-1)\alpha_{k-1}]-(\beta_0+\beta_1+\cdots+\beta_k),\\
c_r &= \frac{1}{r!}[k^r-(\alpha_1+2^r\alpha_2+\cdots+(k-1)^r\alpha_{k-1}]\\
&\quad -\frac{1}{(r-1)!}(\beta_1+2^{r-1}\beta_2+\cdots+k^{r-1}\beta_k \quad (r=2,3,4,\cdots)
\end{aligned} \tag{9.43}$$

若在式(9.40)中选择系数 α_i 和 β_i，满足

$$c_0=c_1=\cdots=c_p=0, c_{p+1}\neq 0.$$

由定义可知此时构造的多步法是 p 阶的，且

$$T_{n+k}=c_{p+1}h^{p+1}y^{(p+1)}(x_n)+O(h^{p+2}). \tag{9.44}$$

其中，$c_{p+1}h^{p+1}y^{(p+1)}(x_n)$称为局部截断误差主项，$c_{p+1}$称为误差常数.

9.4.2 基于数值积分的方法

下面首先讨论基于数值积分的方法来构造形如式(9.40)的线性多步法公式.

将初值问题式(9.2)中的方程在区间$[x_n,x_{n+4}]$上积分

$$y(x_{n+4})-y(x_n)=\int_{x_n}^{x_{n+4}}f(x,y(x))\mathrm{d}x,$$

被积函数用在 x_{n+1},x_{n+2},x_{n+3}处的二次 Lagrange 插值多项式

$$L_2(x)=\sum_{i=0}^{2}f(x_{n+1+i},y(x_{n+1+i}))l_i(x)$$

代替，其中 $l_i(x)$是 Lagrange 插值基函数

$$l_i(x)=\prod_{\substack{j=0\\ j\neq i}}^{2}\frac{x-x_{n+1+j}}{x_{n+1+j}-x_{n+1+j}}.$$

对插值多项式积分得到

$$\int_{x_n}^{x_{n+4}}L_2(x)\mathrm{d}x=\frac{4}{3}h[2f(x_{n+1},y(x_{n+1}))-f(x_{n+2},y(x_{n+2}))+2f(x_{n+3},y(x_{n+3}))].$$

用 y_j 表示 $y(x_j)$的近似值，记 $f_j=f(x_j,y_j)$，从而得到

$$y_{n+4}=y_n+\frac{4h}{3}(2f_{n+1}-f_{n+2}+2f_{n+3}). \tag{9.45}$$

该方法称为米尔尼(Milne)方法.

由局部截断误差定义可知

$$\begin{aligned}
T_{n+4} &= y(x_{n+4})-y(x_n)-\frac{4h}{3}[2y'(x_{n+1})-y'(x_{n+2})+2y'(x_{n+3})]\\
&= \frac{14}{45}h^5y^{(5)}(x_n)+O(h^6),
\end{aligned}$$

故米尔尼方法式(9.45)是4阶方法.

类似地,将方程从 x_n 到 x_{n+2} 积分,得到

$$y(x_{n+2})-y(x_n)=\int_{x_n}^{x_{n+2}} f(x,y(x))\mathrm{d}x. \tag{9.46}$$

被积函数用其在 x_n,x_{n+1},x_{n+2} 处的二次 Lagrange 插值多项式来代替,可以得到

$$y_{n+2}=y_n+\frac{h}{3}(f_n+f_{n+1}+f_{n+2}). \tag{9.47}$$

该方法称为辛普森(Simpson)方法. 事实上,在式(9.46)中对积分项采用 Simpson 求积公式即可得式(9.47). 容易求得其局部截断误差

$$T_{n+2}=-\frac{1}{90}h^5y^{(5)}(x_n)+O(h^6),$$

从而 Simpson 方法是隐式两步四阶方法.

9.4.3 基于 Taylor 展开的方法

考虑形如

$$y_{n+k}=y_{n+k-1}+h\sum_{i=0}^{k}\beta_i f_{n+i} \tag{9.48}$$

的 k 步法,称为阿当姆斯(Adams)方法. $\beta_k=0$ 时为显式方法,$\beta_k\neq 0$ 时为隐式方法. 通常称为 Adams 显式和隐式方法,也称为 Adams—Bashforth 公式和 Adams—Moulton 公式.

这类公式可由方程两端从 x_{n+k-1} 到 x_{n+k} 积分,得到

$$y(x_{n+k})-y(x_{n+k-1})=\int_{x_{n+k-1}}^{x_{n+k}} f(x,y(x))\mathrm{d}x.$$

上式右端积分中被积函数分别用函数在节点 $x_n,x_{n+1},\cdots,x_{n+k-1}$ 和节点 $x_n,x_{n+1},\cdots,x_{n+k-1},x_{n+k}$ 处的 Lagrange 插值多项式代替,完全类似 9.4.2 中的做法,就可以得到 Adams—Bashforth 公式和 Adams—Moulton 公式.

下面以 $k=3$ 为例,讨论运用 Taylor 展开的方法来构造多步法公式.

将式(9.48)与式(9.40)对比(这时 $k=3$),可知 $\alpha_0=\alpha_1=0,\alpha_2=1$. 显然 $c_0=0$. 令 $c_1=c_2=c_3=c_4=0$,由式(9.43)可知

$$\begin{cases}\beta_0+\beta_1+\beta_2+\beta_3=1,\\ 2(\beta_1+2\beta_2+3\beta_3)=5,\\ 3(\beta_1+4\beta_2+9\beta_3)=19,\\ 4(\beta_1+8\beta_2+27\beta_3)=65.\end{cases}$$

若 $\beta_3=0$,则由前面三个方程可以解得

$$\beta_0=\frac{5}{12},\beta_1=-\frac{16}{12},\beta_2=\frac{23}{12}.$$

得到 $k=3$ 的 Adams 显式公式为

$$y_{n+3}=y_{n+2}+\frac{h}{12}(23f_{n+2}-16f_{n+1}+5f_n), \tag{9.49}$$

其局部截断误差为

$$T_{n+3}=\frac{3}{8}h^4y^{(4)}(x_n)+O(h^5),$$

故式(9.49)是三步三阶方法.

若 $\beta_3\neq 0$,则可解得

$$\beta_0=\frac{1}{24},\beta_1=-\frac{5}{24},\beta_2=\frac{19}{24},\beta_3=\frac{3}{8}.$$

得到 $k=3$ 的 Adams 隐式公式为

$$y_{n+3}=y_{n+2}+\frac{h}{24}(9f_{n+3}+19f_{n+2}-5f_{n+1}+f_n), \tag{9.50}$$

其局部截断误差为

$$T_{n+3}=-\frac{19}{720}h^5y^{(5)}(x_n)+O(h^6),$$

故式(9.50)是三步四阶方法.

用类似方法可求得 Adams 显式方法和隐式方法($k=1,2,4$)的公式.

利用基于 Taylor 展开的方法来构造线性多步法比较灵活,它可以构造任意多步法公式.这时就不必拘泥于前面的式(9.43),只要选取 α_i、β_i 使线性多步法的阶尽可能高就行.

例 9.9 解初值问题

$$\begin{cases}y'=f(x,y),\\ y(x_0)=y_0.\end{cases}$$

用显式二步法 $y_{n+1}=\alpha_0y_n+\alpha_1y_{n-1}+h(\beta_0f_n+\beta_1f_{n+1})$,其中 $f_n=f(x_n,y_n)$,$f_{n-1}=f(x_{n-1},y_{n-1})$. 试确定参数 α_0、α_1、β_0、β_1 使方法阶数尽可能高,并求局部截断误差.

解 由局部截断误差定义,并利用 Taylor 公式得到

$$\begin{aligned}T_{n+1}&=y(x_n+h)-\alpha_0y(x_n)-\alpha_1y(x_n-h)-h[\beta_0y'(x_n)+\beta_1y'(x_n-h)]\\&=y(x_n)+hy'(x_n)+\frac{h^2}{2}y''(x_n)+\frac{h^3}{3!}y'''(x_n)+\frac{h^4}{4!}y^{(4)}(x_n)+O(h^5)-\alpha_0y(x_n)\\&\quad-\alpha_1[y(x_n)-hy'(x_n)+\frac{h^2}{2}y''(x_n)-\frac{h^3}{3}y'''(x_n)+\frac{h^4}{4}y^{(4)}(x_n)+O(h^5)]-\beta_0hy'(x_n)\\&\quad-\beta_1h[y'(x_n)-hy''(x_n)+\frac{h^2}{2}y'''(x_n)-\frac{h^3}{3!}y^{(4)}(x_n)+O(h^4)]\\&=(1-\alpha_0-\alpha_1)y(x_n)+(1+\alpha_1-\beta_0-\beta_1)hy'(x_n)+(\frac{1}{2}-\frac{1}{2}\alpha_1+\beta_1)h^2y''(x_n)\end{aligned}$$

$$+\left(\frac{1}{6}+\frac{1}{6}\alpha_1-\frac{1}{2}\beta_1\right)h^3\cdot y'''(x_n)+\left(\frac{1}{24}-\frac{1}{24}\alpha_1+\frac{1}{6}\beta_1\right)h^4y^{(4)}(x_n)+O(h^5).$$

为求参数 α_0、α_1、β_0、β_1 使方法阶数尽可能高，令

$$1-\alpha_0-\alpha_1=0,\qquad 1+\alpha_1-\beta_0-\beta_1=0,$$

$$\frac{1}{2}-\frac{1}{2}\alpha_1+\beta_1=0,\qquad \frac{1}{6}+\frac{1}{6}\alpha_1-\frac{1}{2}\beta_1=0.$$

解得 $\alpha_0=-4,\alpha_1=5,\beta_0=4,\beta_1=2$. 此时公式为三阶，而且局部截断误差为

$$T_{n+1}=\frac{1}{6}h^4y^{(4)}(x_n)+O(h^5),$$

从而所求二步法为

$$y_{n+1}=-4y_n+5y_{n-1}+2h(2f_n+f_{n-1}).$$

例 9.10 证明线性二步法

$$y_{n+2}+(b-1)y_{n+1}-by_n=\frac{h}{4}[(b+3)f_{n+2}+(3b+1)f_n]$$

当 $b\neq-1$ 时方法为二阶，当 $b=-1$ 时方法为三阶.

证明 方法的局部截断误差为

$$T_{n+2}=y(x_{n+2})+(b-1)y(x_{n+1})-by(x_n)-\frac{h}{4}[(b+3)y'(x_{n+2})+(3b+1)y'(x_n)],$$

利用 Taylor 展开得到

$$y(x_{n+2})=y(x_n)+2hy'(x_n)+\frac{(2h)^2}{2}y''(x_n)+\frac{(2h)^3}{3!}y'''(x_n)+\frac{(2h)^4}{4!}y^{(4)}(x_n)+O(h^5),$$

$$y(x_{n+1})=y(x_n)+hy'(x_n)+\frac{h^2}{2}y''(x_n)+\frac{h^3}{3!}y'''(x_n)+\frac{h^4}{4!}y^{(4)}(x_n)+O(h^5),$$

$$y'(x_{n+2})=y'(x_n)+2hy''(x_n)+\frac{(2h)^2}{2}y'''(x_n)+\frac{(2h)^3}{3!}y^{(4)}(x_n)+O(h^4).$$

从而 T_{n+2} 可以表示为

$$\begin{aligned}T_{n+2}=&[1+(b-1)-b]y(x_n)+\left[2+(b-1)-\frac{1}{4}(b+3+3b+1)\right]hy'(x_n)\\&+\left[2+\frac{1}{2}(b-1)-\frac{1}{2}(b+3)\right]h^2y''(x_n)\\&+\left[\frac{4}{3}+\frac{1}{6}(b-1)-\frac{1}{2}(b+3)\right]h^3y'''(x_n)\\&+\left[\frac{2^4}{4!}+\frac{1}{4!}(b-1)-\frac{1}{4}(b+3)\times\frac{8}{3!}\right]h^4y^{(4)}(x_n)+O(h^5)\\=&-\frac{1}{3}(b+1)h^3y'''(x_n)-\frac{7b+9}{24}h^4y^{(4)}(x_n)+O(h^5).\end{aligned}$$

从而当 $b\neq -1$ 时，$T_{n+2}=-\frac{1}{3}(b+1)h^3y'''(x_n)+O(h^4)$，方法为二阶．当 $b=-1$ 时，$T_{n+2}=-\frac{1}{12}h^4y^{(4)}(x_n)+O(h^5)$，方法为三阶．

9.4.4 预测—校正方法

显式多步法计算简单，但是其精度及计算的稳定性没有隐式方法好．隐式多步法一般需采用迭代计算，计算量大．在实际应用中，隐式法一般不单独使用，而是用于改善用显式方法计算得到的近似值．由显式方法给出预测，再用隐式方法校正该预测值，这样得到的算法称为预测—校正方法．

一般情况下，预测公式与校正公式都取同阶的显式方法与隐式方法相匹配．例如用四阶的 Adams 显式方法

$$y_{n+4}=y_{n+3}+\frac{h}{24}(55f_{n+3}-59f_{n+2}+37f_{n+1}-9f_n)$$

做预测，用四阶的 Adams 隐式方法[即式(9.50)]

$$y_{n+4}=y_{n+3}+\frac{h}{24}(9f_{n+4}+19f_{n+3}-5f_{n+2}+f_{n+1})$$

做校正，得到如下预测—校正方法．

预测：
$$\bar{y}_{n+4}=y_{n+3}+\frac{h}{24}[55f(x_{n+3},y_{n+3})-59f(x_{n+2},y_{n+2})+37f(x_{n+1},y_{n+1})-9f(x_n,y_n)],$$

校正：
$$y_{n+4}=y_{n+3}+\frac{h}{24}[9f(x_{n+4},\bar{y}_{n+4})+19f(x_{n+3},y_{n+3})-5f(x_{n+2},y_{n+2})+f(x_{n+1},y_{n+1})],$$

其中，用四阶 R－K 方法提供迭代初值 y_1,y_2,y_3.

9.5 一阶方程组和高阶方程

9.5.1 一阶方程组

对单个方程 $y'=f(x,y)$的数值解法，只要把 y 和 f 理解为向量，即可应用到一阶方程组的情形．

考察一阶方程组

$$\begin{cases}y'_i=f_i(x,y_1,y_2,\cdots,y_N),\\ y_i(x_0)=y_i^0.\end{cases}(i=1,2,\cdots,N)\tag{9.51}$$

采用向量的写法，记

$$\boldsymbol{Y}=(y_1,y_2,\cdots,y_N)^{\mathrm{T}},$$
$$\boldsymbol{Y}^0=(y_1^0,y_2^0,\cdots,y_N^0)^{\mathrm{T}},$$
$$\boldsymbol{F}=(f_1,f_2,\cdots,f_N)^{\mathrm{T}}.$$

方程组式(9.51)的初值问题可表为

$$\begin{cases}\boldsymbol{Y}'=\boldsymbol{F}(x,\boldsymbol{Y}),\\ \boldsymbol{Y}(x_0)=\boldsymbol{Y}^0.\end{cases}$$

求解该问题的四阶 R－K 公式为

$$\boldsymbol{Y}_{n+1}=\boldsymbol{Y}_n+\frac{h}{6}(\boldsymbol{K}_1+2\boldsymbol{K}_2+2\boldsymbol{K}_3+\boldsymbol{K}_4),\tag{9.52}$$

其中

$$\boldsymbol{K}_1=\boldsymbol{F}(x_n,\boldsymbol{Y}_n),$$
$$\boldsymbol{K}_2=\boldsymbol{F}(x_n+\frac{h}{2},\boldsymbol{Y}_n+\frac{h}{2}\boldsymbol{K}_1),$$
$$\boldsymbol{K}_3=\boldsymbol{F}(x_n+\frac{h}{2},\boldsymbol{Y}_n+\frac{h}{2}\boldsymbol{K}_2),$$
$$\boldsymbol{K}_4=\boldsymbol{F}(x_n+h,\boldsymbol{Y}_n+h\boldsymbol{K}_3),$$

注意这里的 $\boldsymbol{K}_i$ 是向量．式(9.52)具体写出来就是

$$y_{i,n+1}=y_{in}+\frac{h}{6}(\boldsymbol{K}_{i1}+2\boldsymbol{K}_{i2}+2\boldsymbol{K}_{i3}+2\boldsymbol{K}_{i4})\quad(i=1,2,\cdots,N),$$

其中

$$\boldsymbol{K}_{i1}=f_i(x_n,y_{1n},y_{2n},\cdots,y_{Nn}),$$
$$\boldsymbol{K}_{i2}=f_i(x_n+\frac{h}{2},y_{1n}+\frac{h}{2}\boldsymbol{K}_{11},y_{2n}+\frac{h}{2}\boldsymbol{K}_{21},\cdots,y_{Nn}+\frac{h}{2}\boldsymbol{K}_{N1}),$$
$$\boldsymbol{K}_{i3}=f_i(x_n+\frac{h}{2},y_{1n}+\frac{h}{2}\boldsymbol{K}_{12},y_{2n}+\frac{h}{2}\boldsymbol{K}_{22},\cdots,y_{Nn}+\frac{h}{2}\boldsymbol{K}_{N2}),$$
$$\boldsymbol{K}_{i4}=f_i(x_n+h,y_{1n}+h\boldsymbol{K}_{13},y_{2n}+h\boldsymbol{K}_{23},\cdots,y_{Nn}+h\boldsymbol{K}_{N3}),$$

这里 y_{in} 表示第 i 个未知函数 $y_i(x)$ 在节点 $x_n=x_0+nh$ 的近似值．

当 $N=2$ 时，是含两个方程的特殊情况

$$\begin{cases}y'=f(x,y,z),\\ z'=g(x,y,z),\\ y(x_0)=y_0,\\ z(x_0)=z_0.\end{cases}$$

其四阶 R－K 公式具有形式

$$\begin{cases}y_{n+1}=y_n+\dfrac{h}{6}(L_1+2L_2+2L_3+L_4),\\ z_{n+1}=z_n+\dfrac{h}{6}(M_1+2M_2+2M_3+M_4).\end{cases}\tag{9.53}$$

其中

$$L_1 = f(x_n, y_n, z_n),$$

$$M_1 = g(x_n, y_n, z_n),$$

$$L_2 = f\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}L_1, z_n + \frac{h}{2}M_1\right),$$

$$M_2 = g\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}L_1, z_n + \frac{h}{2}M_1\right),$$

$$L_3 = f\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}L_2, z_n + \frac{h}{2}M_2\right),$$

$$M_3 = g\left(x_n + \frac{h}{2}, y_n + \frac{h}{2}L_2, z_n + \frac{h}{2}M_2\right),$$

$$L_4 = f(x_n + h, y_n + hL_3, z_n + hM_3),$$

$$M_4 = g(x_n + h, y_n + hL_3, z_n + hM_3).$$

这是单步法，利用节点 x_n 上的值，顺序计算 L_1、M_1、L_2、M_2、L_3、M_3、L_4、M_4. 然后代入式(9.53)，即可求得节点 x_{n+1} 上的值 y_{n+1}, z_{n+1}.

9.5.2 化高阶方程为一阶方程组

对高阶方程的初值问题

$$y^{(m)} = f(x, y, y', \cdots y^{(m-1)}), \tag{9.54}$$

$$y(x_0) = y_0, y'(x_0) = y'_0, \cdots, y^{(m-1)}(x_0) = y_0^{(m-1)}, \tag{9.55}$$

只要引进新的变量

$$y_1 = y, y_2 = y', \cdots, y_m = y^{(m-1)}$$

则式(9.54)化为

$$\begin{cases} y'_1 = y_2, \\ y'_2 = y_3, \\ \cdots\cdots \\ y'_{m-1} = y_m, \\ y'_m = f(x, y_1, y_2, \cdots, y_m). \end{cases} \tag{9.56}$$

初始条件式(9.55)相应的化为

$$y_1(x_0) = y_0, y_2(x_0) = y'_0, \cdots, y_m(x_0) = y_0^{(m-1)}. \tag{9.57}$$

显然问题式(9.54)、式(9.55)和问题式(9.56)、式(9.57)是等价的. 按照 9.5.1 中的做法即可对式(9.56)、式(9.57)进行求解.

9.6 边值问题的数值解法

前面主要介绍了常微分方程初值问题的数值解法，这里初步介绍边值问题的数值解法．边值问题的数值解法主要有有限元方法和有限差分法，下面介绍有限差分法．有限差分法的关键是将求解区间(或区域)离散化并用数值微商对导数进行离散．

考虑两点边值问题

$$\begin{cases} -\dfrac{\mathrm{d}}{\mathrm{d}x}\left(p(x)\dfrac{\mathrm{d}y}{\mathrm{d}x}\right)+r(x)\dfrac{\mathrm{d}y}{\mathrm{d}x}+q(x)y=f(x) \quad (a<x<b), \\ y(a)=\alpha, y(b)=\beta. \end{cases} \tag{9.58}$$

假设 $p(x)$在闭区间$[a,b]$上具有一阶连续导数，$p(x)\geqslant p_{\min}>0$，r、q、f 在闭区间$[a,b]$上连续，α、β是给定的常数．

首先取 $N+1$ 个节点：

$$a=x_0<x_1<\cdots<x_i<\cdots<x_N=b,$$

于是将区间 $I=[a,b]$分成 N 个小区间：

$$I_1=[x_0,x_1], I_2=[x_1,x_2],\cdots, I_N=[x_{N-1},x_N],$$

称为将区间 $I=[a,b]$进行网格剖分．记 $h_i=x_i-x_{i-1}$，$i=1,2,\cdots,N$，称 $h=\max\limits_{1\leqslant i\leqslant N}h_i$ 为网格最大步长．$x_1,x_2,\cdots,x_{N-1}$称为网格内点，$x_0=a$ 和 $x_N=b$ 称为网格界点．

取相邻节点 x_{i-1}和 x_i 的中点 $x_{i-\frac{1}{2}}=\dfrac{1}{2}(x_{i-1}+x_i)$$(i=1,2,\cdots,N)$，称为半整数点．则由节点

$$a=x_0<x_{\frac{1}{2}}<\cdots<x_{i-\frac{1}{2}}<\cdots<x_{N-\frac{1}{2}}<x_N=b$$

又构成区间 $I=[a,b]$的一个网格剖分，称为对偶剖分．

其次利用 Taylor 展开，用差商代替微商，将问题式(9.58)中的方程在内点 x_i 处离散化．将 $y(x_{i+1})$和 $y(x_{i-1})$分别在节点 x_i 处做 Taylor 展开得到

$$y(x_{i+1})=y(x_i+h_{i+1})=y(x_i)+h_{i+1}y'(x_i)+\frac{h_{i+1}^2}{2}y''(x_i)+O(h^3), \tag{9.59}$$

$$y(x_{i-1})=y(x_i-h_i)=y(x_i)-h_iy'(x_i)+\frac{h_i^2}{2}y''(x_i)+O(h^3). \tag{9.60}$$

两式相减得到

$$\frac{y(x_{i+1})-y(x_{i-1})}{h_{i+1}+h_i}=\left[\frac{\mathrm{d}y}{\mathrm{d}x}\right]_i+\frac{h_{i+1}-h_i}{2}\left[\frac{\mathrm{d}^2y}{\mathrm{d}x^2}\right]_i+O(h^2). \tag{9.61}$$

这里记号$[u]_i$ 表示 $u(x_i)$．下面的记号$[u]_{i-\frac{1}{2}}$表示 $u(x_{i-\frac{1}{2}})$．

将 $y(x_i)$和 $y(x_{i-1})$分别在 $x_{i-\frac{1}{2}}$处展开，简单的计算后得到

$$p(x_{i-\frac{1}{2}})\frac{y(x_i)-y(x_{i-1})}{h_i}$$

$$= \left[p\frac{\mathrm{d}y}{\mathrm{d}x}\right]_{i-\frac{1}{2}} + \frac{h_i^2}{24}\left[p\frac{\mathrm{d}^3y}{\mathrm{d}x^3}\right]_{i-\frac{1}{2}} + O(h^3)$$

$$= \left[p\frac{\mathrm{d}y}{\mathrm{d}x}\right]_{i-\frac{1}{2}} + \frac{h_i^2}{24}\left[p\frac{\mathrm{d}^3y}{\mathrm{d}x^3}\right]_{i} + O(h^3). \tag{9.62}$$

完全类似得到

$$p(x_{i+\frac{1}{2}})\frac{y(x_{i+1})-y(x_i)}{h_{i+1}} = \left[p\frac{\mathrm{d}y}{\mathrm{d}x}\right]_{i+\frac{1}{2}} + \frac{h_{i+1}^2}{24}\left[p\frac{\mathrm{d}^3y}{\mathrm{d}x^3}\right]_{i} + O(h^3). \tag{9.63}$$

由式(9.63)减去式(9.62)，并除以$\frac{h_i+h_{i+1}}{2}$，得到

$$\begin{aligned}
&\frac{2}{h_i+h_{i+1}}\left[p(x_{i+\frac{1}{2}})\frac{y(x_{i+1})-y(x_i)}{h_{i+1}} - p(x_{i-\frac{1}{2}})\frac{y(x_i)-y(x_{i-1})}{h_i}\right]\\
&= \frac{2}{h_i+h_{i+1}}\left(\left[p\frac{\mathrm{d}y}{\mathrm{d}x}\right]_{i+\frac{1}{2}} - \left[p\frac{\mathrm{d}y}{\mathrm{d}x}\right]_{i-\frac{1}{2}}\right) + \frac{h_{i+1}-h_i}{12}\left[p\frac{\mathrm{d}^3y}{\mathrm{d}x^3}\right]_i + O(h^2)\\
&= \left[\frac{\mathrm{d}}{\mathrm{d}x}\left(p\frac{\mathrm{d}y}{\mathrm{d}x}\right)\right]_i + \frac{h_{i+1}-h_i}{4}\left[\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left(p\frac{\mathrm{d}y}{\mathrm{d}x}\right)\right]_i + \frac{h_{i+1}-h_i}{12}\left[p\frac{\mathrm{d}^3y}{\mathrm{d}x^3}\right]_i + O(h^2).
\end{aligned} \tag{9.64}$$

令 $p_{i-\frac{1}{2}}=p(x_{i-\frac{1}{2}})$，$r_i=r(x_i)$，$q_i=q(x_i)$，$f_i=f(x_i)$. 由问题式(9.58)可知在节点 x_i 处成立

$$-\left[\frac{\mathrm{d}}{\mathrm{d}x}\left(p\frac{\mathrm{d}y}{\mathrm{d}x}\right)\right]_i + r_i\left[\frac{\mathrm{d}y}{\mathrm{d}x}\right]_i + q_iy(x_i) = f_i. \tag{9.65}$$

由式(9.61)、式(9.64)和式(9.65)得到

$$\begin{aligned}
&-\frac{2}{h_i+h_{i+1}}\left[p_{i+\frac{1}{2}}\frac{y(x_{i+1})-y(x_i)}{h_{i+1}} - p_{i-\frac{1}{2}}\frac{y(x_i)-y(x_{i-1})}{h_i}\right]\\
&+\frac{r_i}{h_{i+1}+h_i}[y(x_{i+1})-y(x_{i-1})] + q_iy(x_i) = f_i + R_i(y)
\end{aligned} \tag{9.66}$$

其中

$$R_i(y) = -(h_{i+1}-h_i)\left(\frac{1}{4}\left[\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left(p\frac{\mathrm{d}y}{\mathrm{d}x}\right)\right]_i + \frac{1}{12}\left[p\frac{\mathrm{d}^3y}{\mathrm{d}x^3}\right]_i - \frac{1}{2}\left[r\frac{\mathrm{d}^2y}{\mathrm{d}x^2}\right]_i\right) + O(h^2).$$

设 y_i 是 $y(x_i)$的近似值，在式(9.66)中舍去 $R_i(y)$，得到逼近边值问题式(9.58)和式(9.59)的差分方程

$$\begin{aligned}
&-\frac{2}{h_i+h_{i+1}}\left[p_{i+\frac{1}{2}}\frac{y_{i+1}-y_i}{h_{i+1}} - p_{i-\frac{1}{2}}\frac{y_i-y_{i-1}}{h_i}\right]\\
&+\frac{r_i}{h_{i+1}+h_i}(y_{i+1}-y_{i-1}) + q_iy_i = f_i \quad (i=1,2,\cdots,N-1),
\end{aligned} \tag{9.67}$$

$$y_0 = \alpha, y_N = \beta. \tag{9.68}$$

差分方程式(9.67)、式(9.68)是 $N-1$ 阶的线性代数方程组，求解方程组就得出解 $u(x_i)$在 x_i 的近似值.

在实际计算中，一般采用等距剖分，即 $h_i \equiv h, i=1,2,\cdots,N$，这时式(9.67)化为更简洁的形式

$$-\frac{1}{h^2}\left[p_{i+\frac{1}{2}}y_{i+1}-(p_{i+\frac{1}{2}}+p_{i-\frac{1}{2}})y_i+p_{i-\frac{1}{2}}y_{i-1}\right]+r_i\frac{y_{i+1}-y_{i-1}}{2h}+q_iy_i=f_i. \quad (9.69)$$

这相当于用一阶中心差商、二阶中心差商分别代替方程中的一阶导数和二阶导数的结果.

习　题

1. 取步长 $h=0.2$，用 Euler 法解初值问题

$$\begin{cases} y'=-y-xy^2 & (0\leqslant x\leqslant 0.6), \\ y(0)=1. \end{cases}$$

2. 用梯形公式解初值问题

$$\begin{cases} y'=8-3y & (1\leqslant x\leqslant 2), \\ y(1)=2. \end{cases}$$

取步长 $h=0.2$，小数点后至少保留 4 位.

3. 用改进欧拉法和梯形法解初值问题

$$\begin{cases} y'=x^2+x-y, \\ y(0)=0. \end{cases}$$

取步长 $h=0.1$，计算到 $x=0.5$，并与准确解 $y=-\mathrm{e}^{-x}+x^2-x+1$ 对比.

4. 利用欧拉方法计算积分

$$\int_0^x \mathrm{e}^{-x^2}\,\mathrm{d}x$$

在点 $x=0.5,0.75,1$ 的近似值.

5. 取步长 $h=0.2$，用经典的四阶 R—K 公式解初值问题

$$\begin{cases} y'=x+y & (0<x<1), \\ y(0)=1. \end{cases}$$

6. 试用数值积分方法直接推导两步法公式

$$y_{n+2}-y_{n+1}=\frac{h}{12}(5f_{n+2}+8f_{n+1}-f_n).$$

7. 构造形如

$$y_{n+1}=\alpha(y_n+y_{n-1})+h(\beta_0 f_n+\beta_1 f_{n-1})$$

的多步法公式，使其为二阶方法，并求出其局部截断误差主项.

8. 证明存在 α 的一个值，使线性多步法

$$y_{n+1}+\alpha(y_n-y_{n-1})-y_{n-2}=\frac{1}{2}(3+\alpha)(f_n+f_{n-1})$$

是四阶的.

9. 建立边值问题

$$\begin{cases} -\dfrac{\mathrm{d}^2 y}{\mathrm{d}x^2} + q(x)y = f(x) \quad (a < x < b), \\ y(a) = \alpha, \\ y'(b) = \beta, \end{cases}$$

的一种差分格式.

参 考 文 献

[1] 姜启源,谢金星,叶俊. 数学模型. 3 版. 北京:高等教育出版社,2007.
[2] 边馥萍,侯文华,梁冯珍. 数学模型方法与算法. 北京:高等教育出版社,2005.
[3] 王树禾. 数学模型选讲. 北京:科学出版社,2008.
[4] 陈恩水,王峰. 数学建模与实验. 北京:科学出版社,2008.
[5] 胡运权,郭耀煌. 运筹学教程. 北京:清华大学出版社,2001.
[6] 李志平. 油气层渗流力学. 北京:石油工业出版社,2001.
[7] 葛家理. 现代油藏渗流力学原理(上册). 北京:石油工业出版社,2003.
[8] 李传亮. 油藏工程原理. 北京:石油工业出版社,2005.
[9] 王鸿勋,张士诚. 水力压裂设计数值计算方法. 北京:石油工业出版社,1998.
[10] 李庆扬,易大义,王能超. 数值分析. 4 版. 北京:清华大学出版社,2005.
[11] 关治,陆金甫. 数值方法. 北京:清华大学出版社,2006.
[12] 同济大学计算数学教研室. 现代数值数学和计算. 上海:同济大学出版社,2004.
[13] 文世鹏,张明. 应用数值分析. 3 版. 北京:石油工业出版社,2005.
[14] 冯康,等. 数值计算方法. 北京:国防工业出版社,1978.
[15] 叶其孝,沈永欢,等. 实用数学手册. 2 版. 北京:科学出版社,2006.
[16] 张韵华,奚梅成,陈效群. 数值计算方法与算法. 北京:科学出版社,2006.
[17] 西南石油大学应用数学教研室. 数值计算方法. 成都:四川科学技术出版社,2007.
[18] 李星. 积分方程. 北京:科学出版社,2008.
[19] 李荣华. 偏微分方程数值解法. 北京:高等教育出版社,2005.
[20] 同济大学计算数学教研室. 数值计算解题方法与同步训练. 上海:同济大学出版社,2002.
[21] 车刚明,等. 数值分析典型例题解析及自测试题. 西安:西北工业大学出版社,2002.
[22] Michael T. Heath. Scientific Computing: An Introductory Survery. 北京:清华大学出版社,2001.

附录A 内　积

在线性代数中，$\boldsymbol{R}^n$ 中任意两个向量 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$ 及 $\boldsymbol{y}=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$ 的内积定义为

$$(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}=x_1y_1+x_2y_2+\cdots+x_ny_n.$$

若将它推广到一般的线性空间 X，则有下面的定义．

定义　设 X 是数域 $\boldsymbol{K}$（$\boldsymbol{R}$ 或 $\boldsymbol{C}$）上的线性空间，对于 $\forall u,v\in X$，存在一个唯一的数 (u,v) 与之对应，且满足

(1) $(u,u)\geqslant 0$，$(u,u)=0$ 当且仅当 $u=0$；

(2) $(u,v)=\overline{(v,u)}$；

(3) $(\alpha u,v)=\alpha(u,v)$；

(4) $(u+v,w)=(u,w)+(v,w)$；

则称 (u,v) 是 X 上元素 u 和 v 的内积，X 称为内积空间．

定义中 $\overline{(v,u)}$ 表示复数 (v,u) 的共轭复数．若 $K=R$，则 $(u,v)=(v,u)$，这时 X 称为实内积空间．对于实内积空间，由定义有

$$(\alpha_1u_1+\alpha_2u_2,\beta_1v_1+\beta_2v_2)=\alpha_1\beta_1(u_1,v_1)+\alpha_1\beta_2(u_1,v_2)+\alpha_2\beta_1(u_2,v_1)+\alpha_2\beta_2(u_2,v_2).$$

若内积空间中两个元素的内积等于零即 $(u,v)=0$，则称 u 和 v 正交．

内积空间 X 上，可以由内积诱导出一个范数，即对于 $\forall u\in X$，记

$$\|u\|=\sqrt{(u,u)}.$$

利用内积的定义及下面即将介绍的 Cauchy－Schwarz 不等式可以验证上述定义满足范数公理．

定理　设 X 是内积空间，对于 $\forall u,v\in X$，有

$$|(u,v)|^2\leqslant(u,u)(v,v),$$

称为柯西—施瓦兹(Cauchy－Schwarz)不等式．

附录B 权 函 数

定义 设$[a,b]$是有限或无限区间，在区间$[a,b]$上非负函数$\rho(x)$满足条件：

(1) $\int_a^b x^k\rho(x)\,\mathrm{d}x$存在且为有限值($k=0,1,2,\cdots$)；

(2)对非负的连续函数$g(x)$，若$\int_a^b g(x)\rho(x)\,\mathrm{d}x=0$，则$g(x)\equiv 0$，就称$\rho(x)$为区间$[a,b]$上的权函数.

例 $X=C[a,b]$，对于$\forall f(x),g(x)\in X$，$\rho(x)$是$[a,b]$上的权函数，则可定义

$$(f,g)=\int_a^b \rho(x)f(x)g(x)\mathrm{d}x,$$

称为函数$f(x)$与$g(x)$在$[a,b]$上带权$\rho(x)$的内积.

若权函数$\rho(x)=1$，这时内积就变成了

$$(f,g)=\int_a^b f(x)g(x)\mathrm{d}x$$

附录 C　正交多项式

定义　若 $\forall f(x),g(x)\in C[a,b]$，$\rho(x)$ 是 $[a,b]$ 上的权函数且满足

$$(f,g)=\int_a^b\rho(x)f(x)g(x)\mathrm{d}x=0,$$

则称 $f(x)$ 与 $g(x)$ 在 $[a,b]$ 上带权 $\rho(x)$ 正交．若函数族 $\varphi_0(x),\varphi_1(x),\cdots,\varphi_n(x),\cdots$ 满足关系

$$(\varphi_j,\varphi_k)=\int_a^b\rho(x)\varphi_j(x)\varphi_k(x)\mathrm{d}x=\begin{cases}0 & (j\neq k),\\ A_k>0 & (j=k),\end{cases}$$

就称 $\{\varphi_k\}$ 是 $[a,b]$ 上带权 $\rho(x)$ 的正交函数系(列)；若 $A_k\equiv1$ 就称之为标准正交函数系(列)．

例　三角函数族 $1,\cos x,\sin x,\cos 2x,\sin 2x,\cdots$，就是区间 $[-\pi,\pi]$ 上的正交函数系(权 $\rho(x)=1$)．

注意到

$$(1,1)=\int_{-\pi}^{\pi}1\mathrm{d}x=2\pi,$$

$$(\sin nx,\sin mx)=\int_{-\pi}^{\pi}\sin nx\sin mx\,\mathrm{d}x=\begin{cases}0 & (m\neq n)\\ \pi & (m=n)\end{cases}\quad(m,n=1,2,\cdots),$$

$$(\cos nx,\cos mx)=\int_{-\pi}^{\pi}\cos nx\cos mx\,\mathrm{d}x=\begin{cases}0 & (m\neq n)\\ \pi & (m=n)\end{cases}\quad(m,n=1,2,\cdots),$$

$$(\cos nx,\sin mx)=\int_{-\pi}^{\pi}\cos nx\sin mx\,\mathrm{d}x=0\quad(n=0,1,2,\cdots;m=1,2,\cdots).$$

定义　若 $\varphi_n(x)$ 是 $[a,b]$ 上首项系数 $a_n\neq0$ 的 n 次多项式，$\rho(x)$ 是 $[a,b]$ 上的权函数，满足

$$\int_a^b\rho(x)\varphi_j(x)\varphi_k(x)\mathrm{d}x=\begin{cases}0 & (j\neq k),\\ A_k>0 & (j=k),\end{cases}$$

就称多项式序列 $\{\varphi_n(x)\}$ 在 $[a,b]$ 上带权 $\rho(x)$ 正交，并称 $\varphi_n(x)$ 是 $[a,b]$ 上带权 $\rho(x)$ 的 n 次正交多项式．

下面给出常见的而又十分重要的几类正交多项式．

C.1　勒让德多项式

在区间 $[-1,1]$ 上，权函数 $\rho(x)\equiv1$ 时的正交多项式称为勒让德(Legendre)多项式，其表达式为

$$P_0(x)=1,P_n(x)=\frac{1}{2^n n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\{(x^2-1)^n\}\quad(n=1,2,\cdots).$$

由于$(x^2-1)^2$是$2n$次多项式，求n阶导数后得

$$P_n(x)=\frac{1}{2^n n!}(2n)(2n-1)\cdots(n+1)x^n+a_{n-1}x^{n-1}+\cdots+a_0.$$

于是得首项x^n的系数$a_n=\frac{(2n)!}{2^n(n!)^2}$. 显然最高项系数为1的勒让德多项式为

$$\widetilde{P}_n(x)=\frac{n!}{(2n)!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\{(x^2-1)^n\}.$$

勒让德多项式有下述几个重要性质：

性质1 正交性

$$(P_n(x),P_m(x))=\int_{-1}^{1}P_n(x)P_m(x)\mathrm{d}(x)=\begin{cases}0 & (m\neq n),\\ \dfrac{2}{2n+1} & (m=n).\end{cases}$$

证明 令$\varphi(x)=(x^2-1)^n$，则$\varphi^{(k)}(\pm1)=0(k=0,1,\cdots,n-1)$且

$$P_n(x)=\frac{1}{2^n n!}\varphi^{(n)}(x).$$

设$Q(x)$是在区间$[-1,1]$上有n阶连续导数的函数，由分部积分知

$$\begin{aligned}\int_{-1}^{1}P_n(x)Q(x)\mathrm{d}x&=\frac{1}{2^n n!}\int_{-1}^{1}Q(x)\varphi^{(n)}(x)\mathrm{d}x\\&=-\frac{1}{2^n n!}\int_{-1}^{1}Q'(x)\varphi^{(n-1)}(x)\mathrm{d}x\\&=\cdots\\&=\frac{(-1)^n}{2^n n!}\int_{-1}^{1}Q^{(n)}(x)\varphi(x)\mathrm{d}x.\end{aligned}$$

下面分两种情况讨论.

(1)若$Q(x)$是次数小于n的多项式，则$Q^{(n)}(x)=0$，故得

$$\int_{-1}^{1}P_m(x)P_n(x)\mathrm{d}x=0\quad(n\neq m).$$

(2)若$Q(x)=P_n(x)=\frac{1}{2^n n!}\varphi^{(n)}(x)=\frac{(2n)!}{2^n(n!)^2}x^n+\cdots$，

$$Q^{(n)}(x)=P_n^{(n)}(x)=\frac{(2n)!}{2^n n!},$$

于是

$$\int_{-1}^{1}P_n^2(x)\mathrm{d}x=\frac{(-1)^n(2n)!}{2^{2n}(n!)^2}\int_{-1}^{1}(x^2-1)^n\mathrm{d}x=\frac{(2n)!}{2^{2n}(n!)^2}\int_{-1}^{1}(1-x^2)^n\mathrm{d}x.$$

由于

$$\int_0^1(1-x^2)^n\mathrm{d}x=\int_0^{\frac{\pi}{2}}\cos^{2n+1}t\mathrm{d}t=\frac{2\cdot4\cdot\cdots\cdot(2n)}{1\cdot3\cdot\cdots\cdot(2n+1)},$$

故

$$\int_{-1}^{1} P_n^2(x)\mathrm{d}x = \frac{2}{2n+1}.$$

性质 2 奇偶性

$$P_n(-x) = (-1)^n P_n(x).$$

由于 $\varphi(x)=(x^2-1)^n$ 是偶次多项式，经过偶次求导仍为偶次多项式，经过奇次求导则为奇次多项式，故 n 为偶数时 $P_n(x)$ 为偶函数，n 为奇数时 $P_n(x)$ 为奇函数.

性质 3 递推关系

考虑 $n+1$ 次多项式 $xP_n(x)$，它可表示为

$$xP_n(x) = a_0P_0(x) + a_1P_1(x) + \cdots + a_{n+1}P_{n+1}(x),$$

两边乘 $P_k(x)$，并从 -1 到 1 积分，得

$$\int_{-1}^{1} xP_n(x)P_k(x)\mathrm{d}x = a_k\int_{-1}^{1} P_k^2(x)\mathrm{d}x$$

当 $k\leqslant n-2$ 时，$xP_k(x)$ 次数小于等于 $n-1$，上式左端积分为 0，故得 $a_k=0$. 当 $k=n$ 时，$xP_n^2(x)$ 为奇函数，左端积分仍为 0，故 $a_n=0$ 于是

$$xP_n(x) = a_{n-1}P_{n-1}(x) + a_{n+1}P_{n+1}(x),$$

其中

$$a_{n-1} = \frac{2n-1}{2}\int_{-1}^{1} xP_n(x)P_{n-1}(x)\mathrm{d}x = \frac{2n-1}{2}\cdot\frac{2n}{4n^2-1} = \frac{n}{2n+1},$$

$$a_{n+1} = \frac{2n+3}{2}\int_{-1}^{1} xP_n(x)P_{n+1}(x)\mathrm{d}x = \frac{2n+3}{2}\cdot\frac{2(n+1)}{(2n+1)(2n+3)} = \frac{n+1}{2n+1}.$$

从而得到以下的递推公式

$$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x) \quad (n=1,2,\cdots).$$

由 $P_0(x)=1$，$P_1(x)=x$，利用上式就可推出

$$P_2(x) = (3x^2-1)/2,$$

$$P_3(x) = (5x^3-3x)/2,$$

$$P_4(x) = (35x^4-30x^2+3)/8,$$

$$P_5(x) = (63x^5-70x^3+15x)/8,$$

$$P_6(x) = (231x^6-315x^4+105x^2-5)/16.$$

C.2 切比雪夫多项式

在区间$[-1,1]$上，权函数 $\rho(x)=\dfrac{1}{\sqrt{1-x^2}}$时的正交多项式就是第一类切比雪夫(Cheby-

shev)多项式，它可表为

$$T_n(x) = \cos(n\arccos x) \quad |x| \leqslant 1.$$

若令 $x=\cos\theta$，则 $T_n(x)=\cos n\theta$ $(0\leqslant\theta\leqslant\pi)$.

切比雪夫多项式有很多重要性质：

性质 1 切比雪夫多项式$\{T_n(x)\}$在区间$[-1,1]$上带权 $\rho(x)=1/\sqrt{1-x^2}$ 正交，且

$$\int_{-1}^{1} \frac{T_n(x)T_m(x)\mathrm{d}x}{\sqrt{1-x^2}} = \begin{cases} 0 & (n \neq m), \\ \dfrac{\pi}{2} & (n = m \neq 0), \\ \pi & (n = m = 0). \end{cases}$$

事实上，令 $x=\cos\theta$，则 $\mathrm{d}x=-\sin\theta\mathrm{d}\theta$，于是

$$\int_{-1}^{1} \frac{T_n(x)T_m(x)\mathrm{d}x}{\sqrt{1-x^2}} = \int_0^{\pi} \cos n\theta \cos m\theta \mathrm{d}\theta = \begin{cases} 0 & (n \neq m), \\ \dfrac{\pi}{2} & (n = m \neq 0), \\ \pi & (n = m = 0). \end{cases}$$

性质 2 递推关系

$$T_0(x) = 1, \quad T_1(x) = x,$$

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \quad (n = 1,2,\cdots).$$

令 $x=\cos\theta$，$T_{n+1}(x)=\cos(n+1)\theta$，由三角恒等式

$$\cos(n+1)\theta + \cos(n-1)\theta = 2\cos\theta\cos n\theta$$

即得．由此可推得

$$T_0(x) = 1,$$

$$T_1(x) = x,$$

$$T_2(x) = 2x^2 - 1,$$

$$T_3(x) = 4x^3 - 3x,$$

$$T_4(x) = 8x^4 - 8x^2 + 1,$$

$$T_5(x) = 16x^5 - 20x^3 + 5x,$$

$$T_6(x) = 32x^6 - 48x^4 + 18x^2 - 1.$$

性质 3 $T_n(-x)=(-1)^nT_n(x)$.

性质 4 $T_n(x)$在区间$[-1,1]$上有 n 个零点 $x_k=\cos\dfrac{2k-1}{2n}\pi, k=1,\cdots,n.$

此外，实际计算中时常要求 x^n 用 $T_0,T_1,\cdots T_n$ 的线性组合表示，其公式为

$$x^n = 2^{1-n}\sum_{k=0}^{\left[\frac{n}{2}\right]}\binom{n}{k}T_{n-2k}(x)$$

这里规定 $T_0=1$. $n=1\sim8$ 的结果如下：

$$1 = T_0,$$

$$x = T_1,$$

$$x^2 = \frac{1}{2}(T_0 + T_2),$$

$$x^3 = \frac{1}{4}(3T_1 + T_3),$$

$$x^4 = \frac{1}{8}(3T_0 + 4T_2 + T_4),$$

$$x^5 = \frac{1}{16}(10T_1 + 5T_3 + T_5),$$

$$x^6 = \frac{1}{32}(10T_0 + 15T_2 + 6T_4 + T_6),$$

$$x^7 = \frac{1}{64}(35T_1 + 21T_3 + 7T_5 + T_7),$$

$$x^8 = \frac{1}{128}(35T_0 + 56T_2 + 28T_4 + 8T_6 + T_8).$$

C.3 其他常用的正交多项式

一般地说，如果区间$[a,b]$及权函数$\rho(x)$不同，则得到的正交多项式也不同．除上述两种最重要的正交多项式外，下面再给出三种较常用的正交多项式．

C.3.1 第二类切比雪夫多项式

在区间$[-1,1]$上带权$\rho(x)=\sqrt{1-x^2}$的正交多项式称为第二类切比雪夫多项式，其表达式为

$$U_n(x) = \frac{\sin[(n+1)\arccos x]}{\sqrt{1-x^2}}.$$

由$x=\cos\theta$可得

$$\int_{-1}^{1} U_n(x)U_m(x)\sqrt{1-x^2}\,\mathrm{d}x = \int_0^{\pi}\sin(n+1)\theta\sin(m+1)\theta\,\mathrm{d}\theta$$

$$= \begin{cases} 0 & (m \neq n), \\ \dfrac{\pi}{2} & (m = n), \end{cases}$$

即$\{U_n(x)\}$是$[-1,1]$上带权$\rho(x)=\sqrt{1-x^2}$的正交多项式族．还可得到递推关系式

$$U_0(x)=1, U_1(x)=2x,$$

$$U_n(x)=2xU_n(x)-U_{n-1}(x) \quad (n=1,2,\cdots).$$

C.3.2 拉盖尔多项式

在区间$[0,\infty]$上带权e^{-x}的正交多项式称为拉盖尔(Laguerre)多项式，其表达式为

$$L_n(x)=e^x\frac{d^n}{dx^n}(x^n e^{-x}).$$

它也具有正交性质

$$\int_0^\infty e^{-x}L_n(x)L_m(x)dx=\begin{cases}0 & (m\neq n),\\ (n!)^2 & (m=n).\end{cases}$$

和递推关系

$$L_0(x)=1, L_1(x)=1-x,$$

$$L_{n+1}(x)=(1+2n-x)L_n(x)-n^2L_{n-1}(x) \quad (n=1,2,\cdots).$$

C.3.3 埃尔米特多项式

在区间$(-\infty,\infty)$上带权e^{-x^2}的正交多项式称为埃尔米特(Hermite)多项式，其表达式为

$$H_n(x)=(-1)^n e^{x^2}\frac{d^n}{dx^n}(e^{-x^2}).$$

它满足正交关系

$$\int_{-\infty}^{=\infty}e^{-x^2}H_m(x)H_n(x)dx=\begin{cases}0 & (m\neq n),\\ 2^n n!\sqrt{\pi} & (m=n),\end{cases}$$

并有递推关系

$$H_0(x)=1, \quad H_1(x)=2x,$$

$$H_{n+1}(x)=2xH_n(x)-2nH_{n-1}(x) \quad (n=1,2,\cdots).$$

[General Information]
书名=工程数学模型及数值计算方法
作者=刘小华编
页数=166
SS号=13519284
DX号=
出版日期=2014.05
出版社=石油工业出版社